

CONFERENCES IN RESEARCH AND PRACTICE IN
INFORMATION TECHNOLOGY

VOLUME 56

USE OF VISION IN HUMAN-COMPUTER
INTERACTION



AUSTRALIAN
COMPUTER
SOCIETY



USE OF VISION IN HUMAN-COMPUTER INTERACTION

Proceedings of the
HCSNet Workshop on the Use of Vision in
Human-Computer Interaction, (VisHCI 2006),
Canberra, Australia, November 2006

Roland Göcke, Antonio Robles-Kelly and Terry Caelli,
Eds.

Volume 56 in the Conferences in Research and Practice in Information Technology Series.
Published by the Australian Computer Society Inc.



Published in association with the ACM Digital Library.

Use of Vision in Human-Computer Interaction. Proceedings of the HCSNet Workshop on the Use of Vision in Human-Computer Interaction, (VisHCI 2006), Canberra, Australia, November 2006

Conferences in Research and Practice in Information Technology, Volume 56.

Copyright ©2006, Australian Computer Society. Reproduction for academic, not-for profit purposes permitted provided the copyright text at the foot of the first page of each paper is included.

Editors:

Roland Göcke,

Antonio Robles-Kelly,

Terry Caelli

National ICT Australia

Locked Bag 8001,

Canberra ACT 2601, Australia

E-mail: roland.goecke@anu.edu.au

Series Editors:

Vladimir Estivill-Castro, Griffith University, Queensland

John F. Roddick, Flinders University, South Australia

Simeon Simoff, University of Technology, Sydney, NSW

crpit@infoeng.flinders.edu.au

Publisher: Australian Computer Society Inc.

PO Box Q534, QVB Post Office

Sydney 1230

New South Wales

Australia.

Conferences in Research and Practice in Information Technology, Volume 56.

ISSN 1445-1336.

ISBN 1-920-68238-4.

Printed, March 2007 by Flinders Press, PO Box 2100, Bedford Park, SA 5042, South Australia.

Cover Design by Modern Planet Design, (08) 8340 1361.

The *Conferences in Research and Practice in Information Technology* series aims to disseminate the results of peer-reviewed research in all areas of Information Technology. Further details can be found at <http://crpit.com/>.

Table of Contents

Proceedings of the HCSNet Workshop on the Use of Vision in Human-Computer Interaction, (VisHCI 2006), Canberra, Australia, November 2006

Preface	vii
Major Sponsor	ix

Keynote Paper

Fast and Accurate Active Appearance Models	3
<i>Iain Matthews</i>	
Audio-Visual Speech Processing: Progress and Challenges	5
<i>Gerasimos Potamianos</i>	
Audio-Visual Technologies for Lecture and Meeting Analysis inside Smart Rooms	7
<i>Gerasimos Potamianos</i>	
Vision in HCI: Embodiment, Multimodality and Information Capacity	9
<i>David Powers</i>	

Full Papers

Vowel recognition of English and German language using Facial movement(SEMG) for Speech control based HCI	13
<i>Sridhar P. Arjunan, Hans Weghorn, Dinesh K. Kumar and Wai C. Yau</i>	
Image-Based Multi-view Scene Analysis using ‘Conexels’	19
<i>Josep R. Casas and Jordi Salvador</i>	
Image Feature Evaluation for Contents-based Image Retrieval	29
<i>Adam Kuffner and Antonio Robles-Kelly</i>	
Observer Annotation of Affective Display and Evaluation of Expressivity: Face vs. Face-and-Body ...	35
<i>Hatice Gunes and Massimo Piccardi</i>	
Face Refinement through a Gradient Descent Alignment Approach	43
<i>Simon Lucey and Iain Matthews</i>	
Learning Active Appearance Models from Image Sequences	51
<i>Jason Saragih and Roland Goecke</i>	
Using Optical Flow for Step Size Initialisation in Hand Tracking by Stochastic Optimisation	61
<i>Desmond Chik</i>	
Hand gestures for HCI using ICA of EMG	67
<i>Ganesh R. Naik, Dinesh Kant Kumar, Vijay Pal Singh and Marimuthu Palaniswam</i>	
Nuisance Free Recognition of Hand Postures Over a Tabletop Display	73
<i>João Carreira and Paulo Peixoto</i>	
Patch-Based Representation of Visual Speech	79
<i>Patrick Lucey and Sridha Sridharan</i>	

Audio-Visual Speaker Verification using Continuous Fused HMMs	87
<i>David Dean, Sridha Sridharan and Tim Wark</i>	
Voiceless Speech Recognition Using Dynamic Visual Speech Features	93
<i>Wai Chee Yau, Dinesh Kant Kumar and Sridhar Poosapadi Arjunan</i>	

Abstracts

Emotions in HCI - An Affective E-Learning System	105
<i>Robin Kaiser and Karina Oertel</i>	
Video to the Rescue	107
<i>Girija Chetty and Michael Wagner</i>	
Author Index	109

Preface

The goal of VisHCI is to provide an Australian-based, international forum for the presentation and discussion of current trends and recent ideas and results from leading national and international scientists to foster scholarly exchange and future collaborations in the human communication sciences. Thus, the VisHCI workshop aims at bringing together researchers, practitioners and students from a number of disciplines related to using vision and visual evidence in human-computer interaction (HCI).

Visual communication - such as hand and body gestures, facial expressions, auditory-visual speech, sign language etc. - is a major communication channel for humans. This, together with the availability of low-cost camera technology has led to an increased use of visual evidence in HCI. As a result, VisHCI has had a good reception by both, the academic community and industry. Thanks to the generous support from HCSNet, registration for VisHCI 2006 was free and a number of travel grants were provided for Australian-based participants to help with the cost of travelling to Canberra. This was of particular importance to encourage student submissions. From the total of these, 80% were student papers. This tendency is reflected in the distribution of accepted papers.

The workshop had strong Recognition, Visual Speech, Hand, and Face recognition tracks with an overall acceptance rate of 55%. On the paper awards, the NICTA Best Paper prize was awarded to Josep R. Casas and Jordi Salvador for their paper entitled "Image-Based Multi-view Scene Analysis using Conexels". The ANU RSISE Best Student Paper Award was shared by Hatice Gunes and Jason Saragih. The two co-winning papers are on the topics of affective computing and facial feature tracking. This highlights the fact that Visual HCI research is multi-disciplinary and both computer vision and HCI research have a strong tradition in Australia.

It is worth noting that the accepted full papers are intended to abide to the Department of Education, Science and Training (DEST) E1 classification for peer-reviewed conference publications. That is, full-length papers underwent a blind, peer review process. Every submitted paper was reviewed by at least two of the programme committee members. The decisions of acceptance for each paper were taken based upon scores provided by the reviewers on presentation, relevance, originality and technical correctness.

Roland Göcke,
Antonio Robles-Kelly,
Terry Caelli
Programme Chairs, VisHCI 2006
November, 2006

Programme Committee

Programme Committee

Adrian Bors (University of York, UK)
Mike Brooks (Adelaide University, Australia)
Horst Bunke (University of Bern, Switzerland)
Denis Burnham (University of Western Sydney, Australia)
Terry Caelli (National ICT Australia / Australian National University, Australia)
Julien Epps (National ICT Australia, Australia)
Roland Göcke (National ICT Australia / Australian National University, Australia)
Hatice Gunes (University of Technology Sydney, Australia)
Edwin Hancock (University of York, UK)
Eun-Jung Holden (University of Western Australia, Australia)
Peter Kovesi (University of Western Australia, Australia)
Simon Lucey (Carnegie Mellon University, USA)
Iain Matthews (Carnegie Mellon University, USA)
Yael Moses (The Interdisciplinary Institute, Israel)
Christian Peter (Fraunhofer IGD, Germany)
Massimo Piccardi (University of Technology Sydney, Australia)
Gerasimos Potamianos (IBM TJ Watson Research Center, USA)
David Powers (Flinders University, Australia)
Eraldo Ribeiro (Florida Institute of Technology, USA)
Antonio Robles-Kelly (National ICT Australia / Australian National University, Australia)
Conrad Sanderson (National ICT Australia, Australia)
Andrea Torsello (Universita' Ca' Foscari di Venezia, Italy)
Jörg Voskamp (Fraunhofer IGD, Germany)

Major Sponsor

The organisers would like to thank the workshop sponsors for their generous support without which it would not have been possible to hold VisHCI 2006 without a registration fee, nor would it have been possible to invite outstanding international keynote speakers or to provide travel grants to participants.

Major Sponsor



Other Sponsors



Research School of Information Sciences and Engineering
<http://csl.rsise.anu.edu.au/>



<http://www.nicta.com.au>



seeingmachines

<http://www.seeingmachines.com/>

KEYNOTE PAPERS

Fast and Accurate Active Appearance Models

Iain Matthews

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: iainm@cs.cmu.edu

Abstract

Active Appearance Models (AAMs) are generative, non-rigid, parametric models of a particular visual phenomena. They are most often applied to faces, which are perhaps the most popular class of deformable objects in computer vision.

There are two elements to using AAMs: first, how to create the model, and second: given a model, how to fit it to a given image? This keynote will give an overview of the standard methods of model construction, then go into detail on the fast, analytical fitting solutions we have developed.

Keywords: Active Appearance Models, non-rigid face tracking, gradient descent image alignment, 2D+3D AAM.

1 Brief Overview of the Keynote

The computer vision community and industry have made great progress on face detection and recognition. The current popular approaches are now mostly based on single image analysis using sliding detection windows and boosted classifiers. They work impressively well for frontal face images. However, there are many more adjuvant application areas for facial analysis if we are able to accurately locate and describe faces in real-time through video sequences. This allows analysis of what a face is "doing", rather than just where it is and who it belongs to. This is a more appealing and difficult problem and is the motivation for our research.

The approach we use is based on Active Appearance Models (Cootes, Edwards & Taylor 2001) (AAMs) but uses an efficient gradient descent fitting algorithm (Matthews & Baker 2004). This approach has led to several extensions to basic algorithm. For example, the 2D+3D algorithm allows us to recover 3D head pose and 3D face shape but can still be fit in real-time (Xiao, Baker, Matthews, & Kanade 2004). We can extend this further to make good use of 3D shape constraints across multiple, simultaneous images (Koterba, Baker, Matthews, Hu, Xiao, Cohn, & Kanade 2005).

The same mathematical framework has also proven useful for automated model construction (Baker, Matthews & Schneider 2004), and robust fitting under occlusion (Gross, Matthews & Baker 2006).

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

The keynote will give an overview of this recent work as well as ongoing extensions and applications, and the current limitations we encounter.

2 Acknowledgements

Many people have contributed to the work that is presented: Simon Baker, Ralph Gross, Jing Xiao, Seth Koterba, Krishnan Ramnath, Changbo Hu, Simon Lucey, Barry Theobald, Takahiro Ishikawa, Junya Inada, Fernando de la Torre, Alvaro Collet, Jeffrey Cohn, Takeo Kanade.

The research was supported in part by DENSO Corporation, the U.S. Department of Defense contract N41756-03-C4024, and the National Institute of Mental Health grant R01 MH51435.

References

- Baker, S., Matthews, I., & Schneider, J. (Oct. 2004), 'Automatic construction of active appearance models as an image coding problem', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 10, pp. 1380–1384.
- Cootes, T.F., Edwards, G.J., & Taylor, C.J. (June 2001), 'Active Appearance Models', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, pp. 681–685.
- Gross, R., Matthews, I., & Baker, S. (June 2006), 'Active appearance models with occlusion', *Image and Vision Computing*, Vol. 24, No. 6, pp. 593–604.
- Koterba, S., Baker, S., Matthews, I., Hu, C., Xiao, J., Cohn, J., & Kanade, T. (2005), 'Multi-view AAM fitting and camera calibration', in *Proc. 10th IEEE International Conference on Computer Vision (ICCV2005)*, Beijing, China, Vol. 1, pp. 511–518.
- Matthews, I., & Baker, S. (Nov. 2004), 'Active appearance models revisited', *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 135–164.
- Xiao, J., Baker, S., Matthews, I., & Kanade, T. (June 2004), 'Real-time combined 2D+3D active appearance models', in *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2004)*, Washington DC, USA, Vol. 2, pp. 535–542.

Audio-Visual Speech Processing: Progress and Challenges

Gerasimos Potamianos

Human Language Technologies Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
Email: gpotam@us.ibm.com

Abstract

This keynote focuses on using visual channel information to improve automatic speech processing for human computer interaction. Two main issues are discussed: the extraction and representation of visual speech, as well as its fusion with traditional acoustic information. The talk mostly considers applying these techniques to automatic speech recognition, however additional areas of interest are also mentioned, for example audio-visual speech detection, enhancement, and synthesis, as well as speaker recognition. The state-of-the-art and remaining challenges in these areas are also discussed.

Keywords: Audio-Visual Speech Processing; Speech Recognition; Speech Enhancement; Speaker Recognition; Speech Synthesis.

Brief Overview of the Keynote

Speech is viewed as an integral part of human-computer interaction (HCI), conveying not only user linguistic information, but also emotion, identity, location, and computer feedback. However, although great progress has been achieved over the past decades, computer processing of speech still lags significantly compared to human performance levels. For example, automatic speech recognition (ASR) lacks robustness to channel mismatch and noise; text-to-speech (TTS) systems continue to lag in naturalness, expressiveness, and, somewhat less, in intelligibility; and typical real-life interaction scenarios, where emotion and non-acoustic cues are used to convey a message, prove insurmountably challenging to traditional computer systems that rely on the audio signal alone. In contrast, humans easily master complex communication tasks by utilizing additional channels of information, most notably the visual sensory channel.

Of central importance to human communication is the visual information present in the face, with the lower face playing an integral role in the production of human speech and of its perception, both being audio-visual in nature. This has motivated significant research over the past quarter century on automatic processing of visual speech and its integration with audio for a number of speech processing applications (Chen 2001). In particular, automatic recognition of visual speech, also known as automatic speechreading, and its fusion with audio-only systems, that gives

rise to audio-visual ASR, has attracted much of this interest (Potamianos et al. 2003). In addition, the need for improved naturalness, expressiveness, and intelligibility of synthesized speech, has steered research work towards augmenting TTS systems by synthesized visual speech (Cosatto and Graf 2001). Further, a number of recently proposed techniques utilize visual-only or joint audio-visual signal processing for speech enhancement, speech activity detection, and source localization, identity recognition from face appearance or visual speech (Chibelushi et al. 2000), and visual recognition and synthesis of human facial emotional expressions. In all cases, the visual modality can significantly improve audio-only systems.

In order to automatically process and incorporate the visual information into the above speech-based HCI technologies, a number of steps are required that are surprisingly similar across them. Central to all technologies is the feature representation of visual speech and its robust extraction. In addition, appropriate integration of the audio and visual representations is required, in order to ensure improved performance of the bimodal systems over audio-only baselines. In a number of technologies, this integration occurs by exploiting audio-visual signal correlation, whereas in others, feature or decision (classifier) fusion techniques are employed. These topics are discussed in detail in this talk, with emphasis on their application to audio-visual ASR. The current state-of-the-art in the area and what is viewed as the remaining challenges to be met are also presented.

Acknowledgements

A number of colleagues at IBM have contributed to the presented work: Stephen M. Chu, Jonathan Connell, Sabine Deligne, Norman Haas, Jing Huang, Giridharan Iyengar, Vit Libal, Etienne Marcheret, Chalapathy Neti, Hariett Nock, Larry Sansone, Andrew W. Senior, and Roberto Sicconi. Furthermore, the following have collaborated with the group through summer internships or postdoctoral fellowships: Ashutosh Garg, Roland Goecke, Guillaume Gravier, Jintao Jiang, Patrick Lucey, and Patricia Scanlon. Finally, collaboration with Petar S. Aleksic and Aggelos K. Katsaggelos (Northwestern U.) on the subject of this talk is also acknowledged.

References

- Chen, T. (2001), "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Processing Mag.*, Vol. 18, No. 1, pp. 9–21.
- Chibelushi, C.C., Deravi, F., & Mason, J.S.D. (2002), "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, Vol. 4, No. 1, pp. 23–37.
- Cosatto, E. & Graf, H.P. (2000), "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, Vol. 2, No. 3, pp. 152–163.
- Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A.W. (2003), "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, Vol. 91, No. 9, pp. 1306–1326.

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Audio-Visual Technologies for Lecture and Meeting Analysis inside Smart Rooms

Gerasimos Potamianos

Human Language Technologies Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
Email: gpotam@us.ibm.com

Abstract

Analysis of lecture meetings recorded inside smart rooms has recently attracted much interest, being the focus of international projects and technology evaluations. In this keynote, we briefly overview one such project, “Computers in the Human Interaction Loop” (CHIL), with emphasis on the perceptual technology components developed. In particular, we focus on person tracking and speech processing technologies, and present the developed IBM systems.

Keywords: Speech Processing; Speech Recognition; Speaker Diarization, Speech Activity Detection; Visual Tracking; Face Detection; Meeting Data.

Brief Overview of the Keynote

Interactive lectures and meetings play significant role in human collaborative activities in the workplace. Not surprisingly, analysis of interaction in these domains has attracted significant interest in the community, being the focus of a number of research efforts and international projects, for example CHIL, AMI, and the U.S. National Institute of Standards and Technology (NIST) Smartspace effort. In these projects, the interaction happens inside smart rooms equipped with multiple audio and visual sensors. Based on the resulting captured data, the goal is to extract higher-level information in order to assist, for example, in lecture meeting indexing, browsing, summarization, and understanding. To achieve this, technology components need to be developed that address basic questions about the “who”, “where”, “what”, and “when” of the interaction.

Addressing these goals is the main aim of the CHIL EU-funded project, run under the technical coordination of the Interactive Systems Laboratories at the University of Karlsruhe, Germany (CHIL project website). In CHIL, computers are reduced to “discreet” observers of human activity through the use of far-field sensors, and are to provide lecture meeting support services to the participants, based on a common architecture that integrates perceptual components. Central to this goal are people tracking and speech processing technologies; in particular *face detection* and three-dimensional (3D) *person tracking*, as well as automatic speech recognition (ASR) or *speech-to-text* (STT), and its complementary technologies, *speech activity detection* (SAD) and *speaker diarization* (SPKR). Significant research effort is be-

ing devoted to developing robust and efficient algorithms to attack these problems. Noticeably, these efforts have been rigorously evaluated in the past few years through project-internal evaluation campaigns, the NIST-sponsored Rich Transcription (RT) Meeting Recognition Evaluation (RT06s evaluation website), and the recent CLEAR (Classification of Events, Activities, and Relationships) campaign (Stiefelhagen et al. 2006).

In this talk, we present a summary of the IBM efforts with respect to the CHIL project, with emphasis on the developed technologies to address face detection, 3D tracking, SAD, SPKR, and STT for the lecture meeting scenario, central to CHIL. Both vision and speech tasks are particular challenging: Speech-wise, due to the presence of multiple speakers with often overlapping speech, a variety of interfering acoustic events (chairs moving, door noise, typing, computer noise, etc.), the strong accents of most speakers and interacting audience members, a high level of spontaneity, hesitations and disfluencies, the technical seminar contents, the relatively small amount of in-domain data, and the use of far-field sensors (Huang et al. 2006); vision-wise, due to low-resolution distant data, people occlusion, and lighting variations (Potamianos and Zhang 2006). Nevertheless, the work reported here shows that addressing these problems in real human interaction scenarios such as CHIL lecture meetings is achievable.

Acknowledgements

A number of colleagues at IBM have been instrumental to the presented work: Stephen M. Chu, Stanley Chen, Jan Curin, Pascal Fleury, Jing Huang, Brian Kingsbury, Jan Kleindienst, Vit Libal, Etienne Marcheret, Chalapaty Neti, Daniel Povey, Thomas Ross, Larry Sansone, Andrew W. Senior, Roberto Sicconi, Olivier Siohan, Alvaro Soneiro, Martin Westphal. Furthermore, Amrith Tyagi and Zhenqiu Zhang have worked on the topics discussed in this work during summer internships at IBM. Support for this work by the European Commission under integrated project CHIL, “Computers in the Human Interaction Loop”, is also acknowledged.

References

- The CHIL Consortium Website, <http://chil.server.de>
Rich Transcription 2006 Spring Meeting Recognition Eval., <http://www.nist.gov/speech/tests/rt/rt2006/spring>
Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., & Soundararajan, P. (2006), “The CLEAR 2006 evaluation,” *Lecture Notes in Computer Science*, Stiefelhagen, R. & Garofolo, J. (eds.), Vol. 4122 (In Press).
Potamianos, G. & Zhang, Z. (2006), “A joint system for single-person 2D-face and 3D-head tracking in CHIL seminars,” *Lecture Notes in Computer Science*, Stiefelhagen, R. & Garofolo, J. (eds.), Vol. 4122 (In Press).
Huang, J., Westphal, M., Chen, S., Siohan, O., Povey, D., Libal, V., Soneiro, A., Schulz, H., Ross, T., and Potamianos, G. (2006), “The IBM Rich Transcription Spring 2006 Speech-To-Text System for Lecture Meetings,” *Machine Learning for Multimodal Interaction* (In Press).

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Vision in HCI: Embodiment, Multimodality and Information Capacity

David Powers

Artificial Intelligence Lab, School of Informatics and Engineering, Flinders University
Adelaide, Australia

Email: david.powers@flinders.edu.au

Abstract

Almost all Human Computer Interfaces involve vision and Pedagogical research encourages the use of multiple modalities including vision. The combination of visual and other modalities, as well as the many submodalities of vision, has both advantages and pitfalls. The work presented here connects psychological research into human cognitive and perceptual processes and limitations, to evaluation and optimization of multimodal HCI.

Keywords: Multimodal interfaces, interface optimization, information capacity.

1 Introduction

Why are educators encouraged to employ multimodal teaching technologies? Why do we like to use Graphical User Interfaces? Why do we need Vision in a Human Computer Interface? And how should we best utilize the various modalities and submodalities?

Most HCI interfaces do involve vision - textual interfaces involve vision both in terms of overt reading but also in terms of orientation within a document or screen. Speech recognition and speech synthesis have their technical issues, but speech has fundamental disadvantages as a sole HCI mechanism versus text, and for programming it is arguably worse - English is not a good programming or representation language, which is why we have designed mathematical and musical notations as well as programming languages.

Similarly speech lacks the persistence and position that text has in relation to other visual elements - that is we can saccade back and forward within a sentence (Huey 1908) or the text, or the program, either consciously or unconsciously, and we retain a 2D or 3D eidetic impression of where we have seen items. The formatting of a text or program, including both left and right indentation, also has a huge impact on how efficiently we can orient in a text and how fast we can read it. Standard typesetting guidelines have been developed over the centuries with a view to optimizing reading speed and orientation.

Graphic User Interfaces add another dimension but there has been a lack of Human Factors analysis in making design decisions, and there is no reason to think the designs we currently have are anything like optimal. Nonetheless, good performance can be achieved with suboptimal interfaces with sufficient training, and a better interface which is demonstrably

more efficient or faster will not necessarily win widespread acceptance given the familiarity and training lock-in phenomenon of which the Qwerty vs Dvorak keyboard is a case in point.

2 Case Studies - Vision Input

2.1 Speech Reading

The classic case of a visual user interface is the use of lip-reading and its fusion with auditory information. Lip reading is good for distinguishing some phonemes that are hard to distinguish aurally, particularly in the face of noise. We will discuss developments relating to finding and tracking facial features as well as fusion of the visual and auditory information in such a way as to guarantee no significant degradation over either alone - viz. no catastrophic fusion (Lewis & Powers 2002, Lewis & Powers 2004). In addition, our research program deals with noise of different kinds, of which lighting and reverb are special cases.

2.2 Situation Awareness

One of the factors that limits the utility of natural language/speech interaction systems is the lack of shared experience/embodiment. Vision is a major part of this and can give the computer a broad view of the world, as well as detail as appropriate.

Further extensions and enhancements in relation to speech reading include the affective interpretation of facial, eye and hand gestures and movements, and the incorporate of muscular (surface EMG/sEMG), ocular (EOG) and brain (EEG) signals. sEMG alone can be used for reasonable lip reading of certain sounds, and the other signals provide correlations with a broader range of linguistic and non-linguistic communication modes and mental states, and represent an integration of AVSR and BCI (Brain Computer Interface). Again careful fusion is necessary to incorporate this information.

3 Case Studies - Vision Output

3.1 Thinking Heads

The above input modalities are complemented by speech synthesis, expression synthesis, dialogue generation and a shared interactive environment, being part of a broad Thinking Head project funded under the ARC/NHMRC Thinking Systems Special Research Initiative.

The full picture is to be exemplified and evaluated in two scenarios - a bill enquiry/complaint scenario and a Second Language (L2) teaching/learning situation. These situations afford opportunity to evaluate appropriateness of computer response and

to characterize user response to different emotional/gestural expressions, including eye gaze and attention tracking. Both scenarios have the opportunity to be enhanced to take into account environmental/situational circumstances/context. In the L2 situation common reference is an essential aspect of the learning situation.

In this case we are not only talking about understanding real auditory and visual input for a Human Head (HH), but modelling and mirroring/simulating the same kind of output with a Thinking Head (TH).

3.2 Simulated Robots

The Robot World (RW) learning situation being imported into the Thinking Head L2 scenario was originally designed for studying Machine Learning of Natural Language and Ontology (L0) and First Language (L1) learning. This system avoids the problems of dealing with real robots and real vision and audition by simulating scripted scenarios and learning or teaching using these scripts.

Grammars, morphologies, ontologies and semantics can all be learned in this L0RW context. The Robot World has its limitations, and a system that is totally simulated based on existing models fails to convince after a point - after all, we are only learning the models we built in. In fact, currently we are working with the CHILDES corpus and building our scenarios around actual sentences and constructs used in child-directed speech. Nonetheless, eventually the robot learners need to see the real world.

3.3 Real Robots

Real robots are able to sense and interact with the real world, and dealing with real robots introduces considerable complexity that takes us away from the human learning and human interface.

Our robotics research has included building a doll that crawls and orients towards a voice, the original version being blind, with a new and rather too heavy head being designed with verging USB cameras and head turning/panning capability. For the new head we also developed an 8-microphone USB array that could be oriented tetrahedrally on the head (ears, mouth and crown) in noise-cancelling 180° pairs for a TH-centred soundfield. The same array can also be worn as a headset for an HH-centred soundfield.

We also use a garbage can on wheels style robot to navigate our building and develop an ontology. Using Wizard of Oz techniques using 802.11 WLAN technology, we have also used it as a building guide. This has a variety of sensors including sonar, an omnidirectional camera. We also use several USB webcams, one of which is used to track our position very precisely. We are also developing a system to read the room numbers (and eventually occupant names and other information). At this stage that is being trained with photos taken from 10 known positions and orientations for each room number, but eventually the image will be taken from the robot's cameras.

3.4 Graphical User Interfaces

The flip side of vision in HCI is the GUI or Graphical User Interface. GUI design has largely neglected human perceptual and cognitive limitations, cognitive load and situation awareness. There has been an implicit assumption that natural is better, and as a corollary, that 3D is better. But this has not been borne out empirically - the converse can be true. Better performance can result from 2D displays in an information retrieval/search context.

We have developed techniques to allow us to display up to 26 simultaneous dimensions in an IR GUI. We are also experimenting with clustering and hyper-space navigation models. But because you can do it doesn't mean you should do it or it is useful to do it.

We therefore have a research focus on understanding the interplay of the linguistic/search dimensions and the visual/graphical dimensions. There are some basic questions about how many dimensions and how many bits of information per dimension people choose to deal with or are capable of dealing with. The work in this area that Miller cited in his Magical Number Seven paper (Miller 1956a), as well as a variety of follow on studies (Miller 1956b), demonstrates that chunking and combination of dimensions can increase the amount of information that can be conveyed to at least 150 distinctions (7 to 8 bits).

In our work we are particularly interested in distinguishing between and controlling for the working memory/cognitive load aspects versus the perceptual aspects, as well as in specifying the optimum matching of application attribute/information dimensions and graphics/display dimensions (Pfitzner, Hobbs & Powers 2003).

We are also evaluating the effectiveness of animation, both as an iconic display dimension and in relation to continuity and situation awareness versus change blindness.

References

- Huey, E.B. (1908), *The Psychology and Pedagogy of Reading*, MIT Press (reprinted 1968).
- Lewis, T.W. & Powers, D.M.W. (2002). 'Audio-Visual Speech Recognition using Red Exclusion and Neural Networks', in *Proc. 25th Australasian Computer Science Conference (ACSC2002)*, Oudshoorn, M.J., ed., Conferences in Research and Practice in Information Technology, Vol. 4, Melbourne, Australia, Australian Computer Society, pp. 149-156.
- Lewis, T.W. & Powers, D.M.W. (2004). 'Sensor Fusion Weighting Measures in Audio-Visual Speech Recognition', in *Proc. 26th Australasian Computer Science Conference (ACSC2004)*, Estivill-Castro, V., ed., Conferences in Research and Practice in Information Technology, Vol. 26, Dunedin, New Zealand, Australian Computer Society, pp. 305-314.
- Miller, G.A. (1956), 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', *Psychology Review*, Vol. 63, pp. 91-97.
- Miller, G.A. (1956), 'Human Memory and the Storage of Information', *IRE Transactions on Information Theory*, Vol. IT-2, No. 3, pp. 129-137.
- Pfitzner, D., Hobbs, V. & Powers, D.M.W. (2003), 'A unified taxonomic framework for information visualization', in *Proc. Australian Symposium on Information Visualization*, Pattison, T. & Thomas, B., eds., Conferences in Research and Practice in Information Technology, Vol. 24, Adelaide, Australia, Australian Computer Society, pp. 57-66.

FULL PAPERS

These papers underwent a blind, peer review process. Every submitted paper was reviewed by at least two members of the programme committee. These papers are intended to abide to the Department of Education, Science and Training (DEST) E1 classification for peer-reviewed conference publications.

Vowel recognition of English and German language using Facial movement(SEMG) for Speech control based HCI

Sridhar P Arjunan¹Hans Weghorn²Dinesh K Kumar¹Wai C Yau¹¹School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia²Information Technology, BA-University of Cooperative Education, Stuttgart, Germany

Email: sridhar_arjunan@ieee.org, weghorn@ba-stuttgart.de, dinesh@rmit.edu.au,

Abstract

This paper examines the use of facial muscle activity (Surface Electromyogram) to recognise speech based commands in English and German language without any audio signals. The system is designed for applications based on speech control for Human Computer Interaction (HCI). The paper presents an effective technique that uses the facial muscle activity of the articulatory muscles and human factors for recognition. The difference in the speed and style of speaking varies between experiments, and this variation appears to be more pronounced when people are speaking a foreign language. To overcome this difficulty, the paper reports measuring the relative activity of the articulatory muscles for recognition of unvoiced vowels of English and German languages. In these investigations, three English vowels and three German vowels were used as recognition variables. The moving root mean square (RMS) of surface electromyogram (SEMG) of four facial muscles is used to segment the signal and to identify the start and end of a silently spoken utterance. The relative muscle activity is computed by integrating and normalising the RMS values of the signals between the detected start and end markers. The output vector of this is classified using a back propagation neural network to identify the voiceless speech. The data is also tested using K means clustering technique to determine the linearity of separation of the data. The experimental results show that this technique gives high recognition rate when used for each of the participants for both of the languages. The investigations also show that the system is easy to train for a new user. The visual inspection of the plot of the experimental data suggests the formation of clusters. The results suggest that such a system is reliable for simple vowel based commands for human computer interface when it is trained for the user, who can speak one or more languages and for the people who have speech disability.

Keywords: Surface Electromyogram, Speech control, HCI, ANN.

1 Introduction

Research and development of new human computer interaction (HCI) techniques that enhance the flexibility and reliability for the user are important. Research on new methods of computer control has fo-

cused on three human factors of body functions: speech, bioelectrical activity and facial expressions. The expression of emotions plays an important part in human interaction. Most of the facial movements result from either speech or the display of emotions; each of these has its own complexity (Ursula 1998)

Speech operated systems have the advantage that these provide the user with flexibility, and can be considered for any applications where natural language may be used. Such systems utilise a natural ability of the user. Such systems have the potential for making computer control effortless and natural. Further, due to the very dense information that can be coded in speech, speech based human computer interaction (HCI) can provide richness comparable to human to human interaction.

In recent years, significant progress has been made in advancing speech recognition technology, making speech an effective modality in both telephony and multimodal human-machine interaction. Speech recognition systems have been built and deployed for numerous applications. The technology is not only improving at a steady pace, but is also becoming increasingly usable and useful. However, speech recognition technology has three major shortcomings; (i) it is not suitable in noisy environments such as a vehicle or a factory, (ii) it is not suitable for people with speech impairment disability, such as people after a stroke attack, and (iii) it is not suitable for giving discrete commands when there may be other people in the vicinity. This paper reports research to overcome these shortcomings, with the intent to develop a system that would identify the verbal command from the user without the need for the user to speak the command. The possible user of such systems would be people with disability, workers in noisy environments, and members of the defence forces.

When we speak in noisy environments, or with people with hearing loss, the lip and facial movements often compensate the lack of quality audio. The identification of the speech with lip movement can be achieved using visual sensing, or sensing of the movement and shape using mechanical sensors (Manabe 2003) or by relating the movement and shape to the muscle activity (Chan 2002, Kumar 2004). Each of these techniques has strengths and limitations. The video based technique is computationally expensive, requires a camera monitoring the lips and fixed to the user's head, and is sensitive to lighting conditions. The sensor based technique has the obvious disadvantage that it requires the user to have sensors fixed to the face, making the system not user friendly. The muscle monitoring systems have limitations of low reliability. This paper reports the use of recording muscle activity of the facial muscles to determine the unspoken command from the user.

Earlier work reported by the authors have demonstrated the use of multi-channel surface electromy-

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

gram (SEMG) to identify the unspoken vowel based on the normalized integral values of SEMG during the utterance. The main common concern with such systems is the difficulty to work across people of different backgrounds and the main challenge is the ability of such a system to work for people of different native languages. Earlier work by the authors had tested the system for native Australian English speakers. This paper compares the error in classification of the unspoken English and German vowels by a group of German native speakers.

2 THEORY

The purpose of this research is to classify the surface recordings of the facial muscle activity with speech. For this analysis, the first step is to determine the role of the facial muscles in the production of speech. There are number of major speech production models that describe the mechanisms of speech productions. It is important to identify the anatomical details of speech production for analysing the shape of the mouth and the muscle activity with speech.

2.1 Articulatory phonetics

Articulatory phonetics considers the anatomical detail of the production speech sounds. This requires the description of speech sounds in terms of the position of the vocal organs. For this purpose, it is convenient to divide the speech sounds into vowels and consonants. The consonants are relatively easy to define in terms of the shape and position of the vocal organs, but the vowels are less well defined and this may be explained because the tongue typically never touches another organ when making a vowel (Thomas 1986). When considering the speech articulation, the shapes of the mouth during speaking vowels remain constant while during consonants the shapes of the mouth changes. The vowel is stationary, while the consonant is non-stationary.

2.2 Face movement related to speech

The face can communicate a variety of information including subjective emotion, communitive intent, and cognitive appraisal. The facial musculature is a three dimensional assembly of small, pseudo-independently controlled muscular lips performing a variety of complex orofacial functions such as speech, mastication, swallowing and mediation of motion (Lapatki 2003). The parameterization used in speech is usually in terms of phonemes. A phoneme is a particular position of the mouth during a sound emission, and corresponds with specific sound properties. These phonemes in turn control the lower level parameters for the actual deformations. The required shape of the mouth and lips for the utterance of the phonemes is achieved by the controlled contraction of the facial muscles that is a result of the activity from the nervous system (Thomas 1986).

Surface electromyogram (SEMG) is the non-invasive recording of the muscle activity. It can be recorded from the surface using electrodes that are stuck to the skin and located close to the muscle to be studied. SEMG is a gross indicator of the muscle activity and is used to identify force of muscle contraction, associated movement and posture (Basmajian 1985). Using an SEMG based system, Chan et al (Chan 2002) demonstrated that the presence of speech information in facial myoelectric signals. Kumar et al (Kumar 2004) have demonstrated the use of SEMG to identify the unspoken sounds under controlled conditions. There are number of

challenges associated with the classification of muscle activity with respect to the associated movement and posture, such as the sensitivity of the location of electrodes, inter user variations, sensitivity of the system to variations in intrinsic factors such as skin conductance, and to external factors such as temperature, and electrode conditions. Veldhuizen et al (Veldhuizen 2003) demonstrated the variation of facial EMG during a single day and has shown facial SEMG activity decreased during the workday and increased again in the evening.

One difficulty with speech identification using facial movement and shape is the temporal variation when the user is speaking complex time varying sounds. With the intra and inter subject variation in the speed of speaking, and the length of each sound, it is difficult to determine a suitable window, and when the properties of the signal are time varying, this makes identifying suitable features for classification less robust. The other difficulties also arise from the need for segmentation and the identification of the start and end of movement if the movement is complex. While each of these challenges are important, as a first step, this paper has considered the use of vowel based verbal commands only, where there is no change in the sound producing apparatus, the mouth cavity and the lips, and the nasal sounds can largely be ignored. Such a system would have limited vocabulary, and would not be very natural, but would be an important step in the evolution. In such a system, using moving RMS threshold, the temporal location of each activity can be identified. By having a stationary set of parameters defining the muscle activity for each spoken event, this also makes the system have very compact set of features, making it suitable for real time classification.

2.3 Facial muscles for speech

When using facial SEMG to determine the shape of the lips and the mouth, there is the issue of the choice of the muscles and the corresponding location of the electrodes. Face structure is more complex than the limbs, with large number of muscles with overlaps. It is thus difficult to identify the specific muscles that are responsible for specific facial actions and shapes. There is also the difficulty of cross talk due to the overlap between the different muscles. This is made more complex due to the temporal variation in the activation and deactivation of the different muscles. The use of integral of the RMS of SEMG is useful in overcoming the issues of cross talk and the temporal difference between the activation of the different muscles that may be close to one set of electrodes. Due to the unknown aspect of the muscle groups that are activated to produce a sound, statistical distance based cluster analysis and back-propagation neural network has been used for classifying the integral of the RMS of the SEMG recordings. It is impractical to consider the entire facial muscles and record their electrical activity. In this study, only four facial muscles have been selected; The *Zygomaticus Major* arises from the front surface of the zygomatic bone and merges with the muscles at the corner of the mouth. The *Depressor anguli oris* originates from the mandible and inserts skin at an angle of mouth and pulls corner of mouth downward. The *Masseter* originates from maxilla and zygomatic arch and inserts to ramus of mandible to elevate and protrude, assists in side-to-side movements mandible. The *Mentalis* originates from the mandible and inserts into the skin of the chin to elevate and protrude lower lip, pull skin into a pout (Fridlund 1986). The location of these muscles are shown in Figure 1.

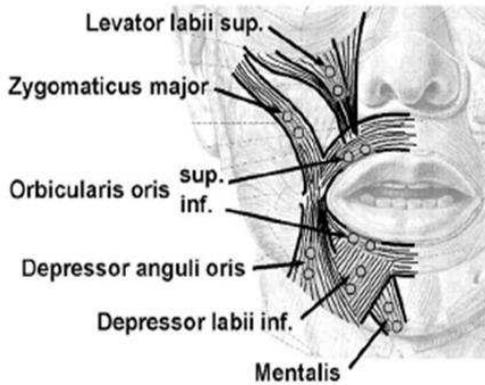


Figure 1: Topographical location of facial muscles [Source: (Lapatki 2003)]

2.4 Features of SEMG

SEMG is a complex and non-stationary signal. The strength of SEMG is a good measure of the strength of contraction of the muscle, and can be related to the movement and posture of the corresponding part of the body. The most commonly used feature to identify the strength of contraction of a muscle is the root mean square (RMS). RMS of SEMG is related to the number of active muscle fibres and the rate of activation, and is a good measure of the strength of the muscle activation, and thus the strength of the force of muscle contraction.

The preliminary study by Chan et al. has demonstrated the presence of speech information in facial EMG (Chan 2002). The timing of the activation of different groups of muscles is a central issue to identify the movement and shape of the mouth and lips. The issue regarding the use of SEMG to identify speech is the large variability of SEMG activity pattern associated with a phoneme of speech. A difference in the amount of motor unit activity was observed in one and the same muscle when different words like p, b were spoken in the same context (Basmajian 1985).

The vowels correspond to stationary muscle activity, the muscle activity pre and post the vowel is non-stationary. While it is relatively simple to identify the start and the end of the muscle activity related to the vowel, the muscle activity at the start and the end may often be much larger than the activity during the section when the mouth cavity shape is being kept constant, corresponding to the vowel. To overcome this issue, this research recommends the use of the integration of the RMS of SEMG from the start till the end of the utterance of the vowel. The temporal location of the start and the end of the activity is identifiable using moving window RMS.

Another shortcoming of the use of strength of SEMG is that it is dependent on the absolute of the magnitude of the recording, which can have large inter experimental variation. To overcome this shortcoming, this paper reports the use of ratios of the area under the curve of SEMG from the different muscles. By taking the ratio rather than the absolute value, the difficulty due the variation of the magnitude of SEMG between different experiments is overcome.

3 Methodology

Experiments were conducted to evaluate the performance of the proposed speech recognition from facial EMG for different languages, German and English. The experiments were approved by the Human Experiments Ethics Committee of the University. Controlled experiments were conducted where the par-

Vowels					
Short	a	e	i	o	u
	[a]	[ɛ]	[ɪ]	[ɔ]	[ʊ]
Long	a/aa/ah	e/ee/eh	i/ih/ie	o/oo/oh	u/uh
	[a:]	[e:]	[i:]	[o:]	[u:]

Figure 2: Pronunciation of German vowels

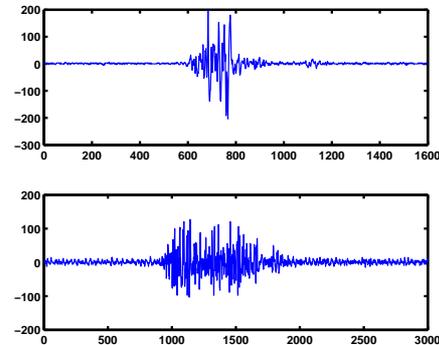


Figure 3: Raw SEMG signal recorded during different experiments

ticipant was asked to speak while their SEMG was recorded. The SEMG recordings were visually observed, and all recordings with any artifacts - typically due to loose electrodes or movement - were discarded. During these recordings, the subjects spoke three selected English vowels (/a/, /e/, /u/) and three selected German vowels (/a/, /i/, /u/). Each vowel was spoken separately such that there was a clear start and end of its utterance. The experiment was repeated ten times for each language. A suitable resting time was given between each experiment. The participants were asked to vary their speaking speed and style to obtain a wide based training set. The pronunciation of German vowels (Ager 2006) is shown in Figure 2.

3.1 EMG Recording and Processing

In previous investigation, three male volunteers speaking English participated and in the present investigation, two male and one female volunteers participated in the experiments. All the participants in this experiment were native speakers of German with English as their second language. Four-channel facial SEMG was recorded using the recommended recording guidelines (Fridlund 1986).

A four channel, portable, continuous recording MEGAWIN instrument (from Mega Electronics, Finland) was used for this purpose. Raw signal was recorded at a rate of 2000 samples/second. Ag/AgCl electrodes (AMBU Blue sensors from MEDICOTEST, Denmark) were mounted on appropriate locations close to the selected facial muscles, which were the right side *Zygomaticus Major*, *Masseter* & *Mentalis* and left side *Depressor anguli oris*. The inter electrode distance was kept constant at 1cm for all the channels and experiments. The recordings were visually observed, and all recordings with any artifacts were discarded. Figure 3 shows the raw SEMG signal recorded during different experiments by changing the speed and style of utterance, plotted as a function of time (sample number).

The first step in the analysis of the data required identifying the temporal location of the muscle activity. Moving root mean square (MRMS) of the recorded signal with a threshold of 1 sigma of the

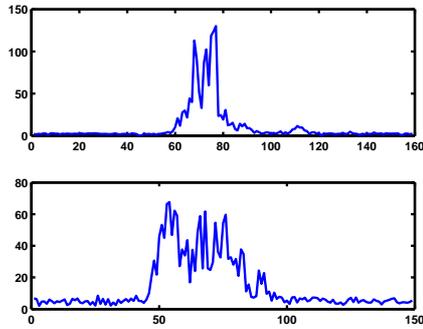


Figure 4: RMS plot of the recorded EMG signals

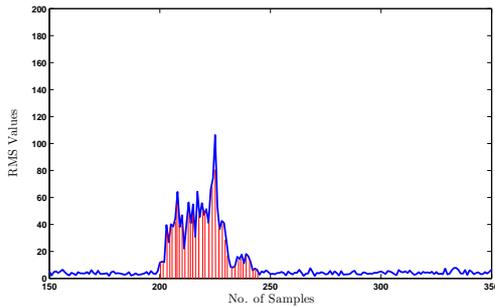


Figure 5: An example of the computation of the integral of RMS of SEMG

signal was applied for windowing and identifying the start and the end of the active period (David 1997). Window size of 20 samples corresponds to 10 msec and was used as the size of the window for computing the MRMS. The start and the end of the muscle activity were also confirmed visually. Figure 4 shows the plot of RMS values of the different recorded EMG signals. The next step is to parameterise the SEMG for classification of the data. To overcome the difference between the speed of utterance during different experiments, and difference between different experiments in the absolute magnitude of the recordings, the data was integrated and normalised. MRMS of the envelope of SEMG between the start and the end of the muscle activity was integrated for each of the channels. This provided a four long vector corresponding to the overall activity of the four channels for each vowel utterance. This data was normalised with respect to channel 1 by computing a ratio of integrated MRMS of each channel with respect to channel 1. This ratio is indicative for the relative strength of contraction of the different muscles and reduces the impact of inter-experiment variations. The outcome of this step was a vector of length three corresponding to each utterance. Figure 5 is an example of the computation of the integral of RMS of SEMG.

For computing the integral of RMS of SEMG, Durand's rule (Eric 2006) was used, because it produces approximations that are more accurate and a straightforward family of numerical integration techniques. A simplified block diagram of methodology shown in Figure 6, explains the process of the analysis.

3.2 Classification of SEMG Data

Parameterization of SEMG data results in a vector with three measures for each utterance. The first step in classification of data was to determine if this data was separable. After confirming this, the next step

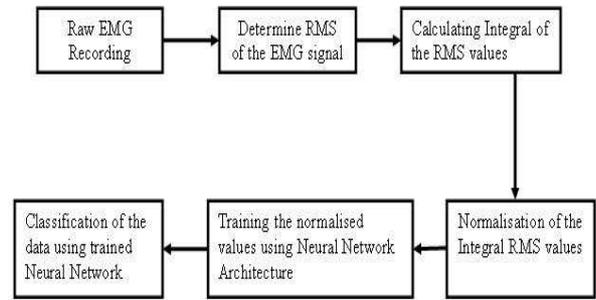


Figure 6: A simplified block diagram of methodology

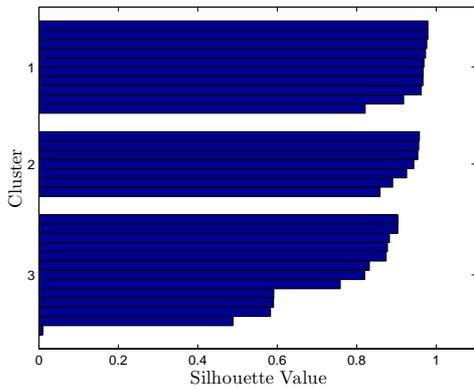
undertaken was to determine whether the data is linearly separable. To determine whether the data is separable, supervised neural network approach was used. The advantage of using such a neural network is that neural networks can be applied without the assumption for linear separation of the data. For this purpose, the data from the ten experiments for each participant was divided into two equal groups - training and test data. This was repeated for English and German language separately. An over-sized neural network was used to ensure identifying the separation of the data.

The ANN consisted of two hidden layers with 20 nodes in both layers. Sigmoid function was used as the threshold decision. ANN was trained with gradient descent algorithm using momentum with a learning rate of 0.05 to reduce likelihood of local minima. Finally, the trained ANN was used to classify the test data. This entire process was repeated for each of the participants. The performance of these integral RMS values was evaluated in this experiment by comparing the accuracy in the classification during testing. The accuracy was computed based on the percentage of correctly classified data points to the total number of data points in the class.

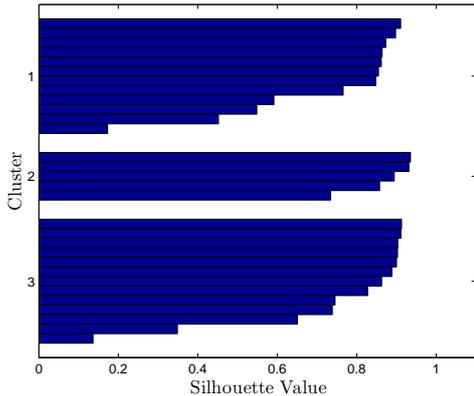
The next step in the classification of this data was to test, whether the data was linearly separable. Taking advantage of the three dimension in the data, three axis plot was produced. In this, data points representing each vowel were given a specific colour and distinct symbol for visual inspection. Figure 8 and Figure 9 show examples of such a plot, for each of the investigated languages. The K-means clustering technique was performed to test the data for linear separability. To get an idea of how well-separated the resulting clusters are, a silhouette plot was made using the cluster indices output from k-means. The silhouette plot in Figure 7 displays a measure of closeness of each point in one cluster is to points in the neighbouring clusters. Unsupervised clustering does not demonstrate linear separability of the data with no temporal information and with a prior knowledge of the targets against the inputs. Supervised back-propagation neural network is the most convenient to use as a classifier, the authors are aware that such a classifier may in some cases be sub-optimum. The advantage of ANN approach is that ANN is easy to be trained by a user to configure the system for the individual.

4 RESULTS AND OBSERVATIONS

The linear separation of normalised integral RMS values of different vowels was tested using three dimensional plot and silhouette plot. It is observable from the 3-D plots in Figure 8 and Figure 9, that there appears distinct clustering of the data based on the vowel uttered for *both* languages. This is also verified using k-means Silhouette plot (Figure 7), it is clear



(a)



(b)

Figure 7: Silhouette plot of the normalised IRMS values (a) English Vowels (b) German Vowels

Table 1: Classification results for different participants uttering English vowels

Vowel	Correctly Classified Vowels		
	Participant 1	Participant 2	Participant 3
/a/	3(60%)	4(80%)	4(80%)
/e/	4(80%)	4(80%)	4(80%)
/u/	5(100%)	5(100%)	5(100%)

that most points have a large silhouette value, indicating that the clusters are separated from each other and it suggests that there exists linear separation of the data. The average silhouette values for English vowels and German vowels are 0.7634 and 0.8441 respectively. This shows that the linear separation of data is stronger in German vowels (native language of the speaker) than English vowels (foreign language).

The results of testing the ANN on the test data using weight matrix generated during training are tabulated in Table 1 for English vowels and Table 2 for German vowels. These results indicate an overall average accuracy of 86%, where it is noted that the overall classification of the integral RMS values of the EMG signal yields better recognition rate of vowels for 3 different participants, when it is trained individually.

The results indicate that this technique can be used for the classification of vowels for the native and foreign language, in this case, English and German. This suggests that the system is able to identify the differences between the styles of speaking of different people at different times for different languages.

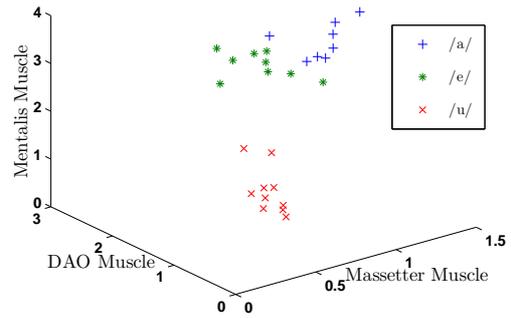


Figure 8: 3-D plot of the normalised IRMS values of English vowels

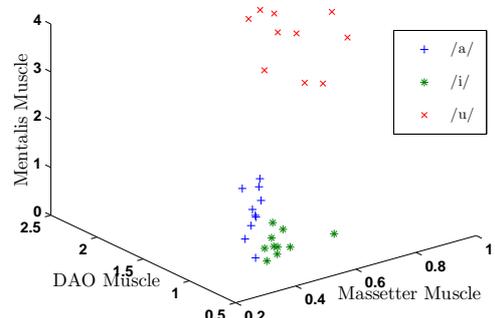


Figure 9: 3-D plot of the normalised IRMS values of German vowels

5 DISCUSSION

The results indicate that the proposed method that uses activities of facial muscles for identifying silently spoken vowels is technically feasible from the view point of error in identification. The investigation reveals the suitability of the system for English and German, and this suggests that the system is feasible when used for people speaking their own native language as well as a foreign language. The results also indicate that the system is not disturbed by the variation in the speed of utterance. The recognition accuracy is high, when it is trained and tested for a dedicate user. Hence, such a system could be used by any individual user as a reliable human computer interface (HCI). This method has only been tested for limited vowels. Vowels were the first to be considered because the muscle contraction during the utterance of vowels remains stationary. The promising results obtained in the experiment indicate that this approach based on the facial muscles movement represents a suitable, reliable method for classifying vowels.

Table 2: Classification results for different participants uttering German vowels

Vowel	Correctly Classified Vowels		
	Participant 1	Participant 2	Participant 3
/a/	4(80%)	4(80%)	4(80%)
/i/	5(100%)	4(80%)	4(80%)
/u/	5(100%)	5(100%)	5(100%)

els of single user without regard to the speaking speed and style in different times for different languages. It should be pointed that this method at this stage is not being designed to provide the flexibility of regular conversation language, but for a limited dictionary only, which is appropriate for simple voice control systems. The results furthermore suggest that such a system is suitable and reliable for simple commands for human computer interface when it is trained for the user. This method has to be enhanced for large set of data with many subjects in future.

6 Conclusion

This paper describes a silent vowel based speech identification approach that is based on measuring the facial muscle contraction using non-invasive SEMG. The experiments indicate that the system is easy to train for a new user and is suitable for two languages - English and German. Application of this include e.g. removal of any disambiguity caused by the acoustic noise for human computer interface or computer based speech control and analysis. The presented investigation focused on classifying English and German vowels, because pronunciation of vowels results in stationary muscle contraction as compared to consonants. The normalised integral RMS values of the facial EMG signals are used for analysis, and classification of these values is performed by ANN. The results indicate that the system is reliable when trained for the individual user. The system has been tested with a very small set of phonemes, where the system has been successful. A broad variety of applications could benefit from this technology: One possible application for such a system is for disabled user to give simple commands to a machine which is a good and typical application of HCI. Future possibilities include applications for telephony, defence problems and improvement of speech-based computer control in noisy environments.

7 Acknowledgments

We would like to thank for the financial support of the Landesstiftung Baden-Württemberg GmbH in Stuttgart, Germany, who co-financed the visiting research placement of the main author of this paper. The supplied funding by the Baden-Württemberg Stipendium enabled that this particular investigation on multi-language voice control could be performed in an efficient manner.

References

- Basmajian, J.V. & DeLuca, C.J. (1985), *Muscles Alive; Their Functions Revealed by Electromyography*, Fifth edition, Williams & Wilkins, Baltimore.
- Thomas W. Parsons. (1986), *Voice and speech processing*, McGraw-Hill, First edition, New York.
- David Freedman, Robert Pisani & Roger Purves. (1997), *Statistics*, Norton College Books, Third Edition, New York.
- Chan, D.C., Englehart, K., Hudgins, B. & Lovely, D. F. (2002), 'A multi-expert speech recognition system using acoustic and myoelectric signals', in *Proceedings 24th Annual Conference and the Annual Fall Meeting of the [Biomedical Engineering Society] EMBS/BMES Conference*, Ottawa, Canada, Vol. 1, pp. 72-73.
- Kumar, S., Kumar, D.K., Alemu, M. & Burry, M. (2004), 'EMG based voice recognition', in *Proceedings of Sensor Networks and Information Processing*, Melbourne, Australia.
- Manabe, H., Hiraiwa, A. & Sugimura, T. (2003), 'Unvoiced speech recognition using SEMG - Mime Speech Recognition', *ACM Conference on Human Factors in Computing Systems*, Ft.Lauderdale, Florida, USA, pp. 794-795.
- Veldhuizen, I.J.T., Gaillard, A.W.K. & de Vries, J. (2003), 'The influence of mental fatigue on facial EMG activity during a simulated workday', *In Journal of Biological Psychology*, Vol. 63, No. 1, pp. 59-78.
- Fridlund, A.J., & Cacioppo, J.T. (1986), 'Guidelines for Human Electromyographic research', *In Journal of Psychophysiology*, Vol. 23, No. 4, pp. 567-589.
- Lapatki, G., Stegeman, D. F. & Jonas, I. E. (2003), 'A surface EMG electrode for the simultaneous observation of multiple facial muscles', *In Journal of Neuroscience Methods*, Vol. 123, No. 2, pp. 117-128.
- Ursula, H., Pierre, P. & Sylvie, B. (1998), 'Facial Reactions to Emotional Facial Expressions: Affect or Cognition?', *Cognition and Emotion*, Vol. 12, No. 4.
- Eric W. Weisstein (2006), Durand's Rule, From MathWorld- A Wolfram Web Resource, <http://mathworld.wolfram.com/DurandsRule.html>. Accessed August 2006.
- Ager, S. (2006), Information about German language, <http://www.omniglot.com/writing/german.htm>. Accessed August 2006.

Image-Based Multi-view Scene Analysis using ‘Conexels’

Josep R. Casas Jordi Salvador

Image Processing Group
UPC – Technical University of Catalonia
Barcelona, Spain

Email: {josep,aljsal}@gps.tsc.upc.edu

Abstract

Multi-camera environments allow constructing volumetric models of the scene to improve the analysis performance of computer vision algorithms (e.g. disambiguating occlusion). When representing volumetric results of image-based multi-camera analysis, a direct approach is to scan the 3D space with regular voxels. Regular voxelization is good at high spatial resolutions for applications such as volume visualization and rendering of synthetic scenes generated by geometric models, or to represent data resulting from direct 3D data capture (e.g. MRI). However, regular voxelization shows a number of drawbacks for visual scene analysis, where direct measurements on 3D voxels are not usually available. In this case, voxel values are computed rather as a result of the analysis on ‘projected’ image data.

In this paper, we first provide some statistics to show how voxels project to ‘unbalanced’ sets of image data in common multi-view analysis settings. Then, we propose a 3D geometry for multi-view scene analysis providing a better balance in terms of the number of pixels used to analyse each elementary volumetric unit. The proposed geometry is non-regular in 3D space, but becomes regular once projected onto camera images, adapting the sampling to the images. The aim is to better exploit multi-view image data by balancing its usage across multiple cameras instead of focusing in regular sampling of 3D space, from which we do not have direct measurements. An efficient recursive algorithm using the proposed geometry is outlined. Experimental results reflect better balance and higher accuracy for multi-view analysis than regular voxelization with equivalent restrictions.

Keywords: Multi-view analysis, volume voxelization, epipolar geometry.

This work has been supported by the EU through IP CHIL IST-2004-506909 and by the Spanish Government through TEC2004-01914.

Copyright © 2006, Australian Computer Society, Inc. This paper appeared at the *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

1 Introduction

The ever decreasing cost of acquisition devices and computing capabilities are making multi-camera settings increasingly common for visual analysis in controlled environments. Multi-view image analysis exploits similarity and disparity in images provided by multiple cameras observing the scene. This offers promising advantages compared to single camera analysis:

- Multi-view analysis algorithms benefit from 3D cues. Non-redundant information extracted from multiple cameras disambiguates occlusions and augments the available visual information with projections from otherwise occluded parts of the scene
- Multi-view image analysis may also yield additional robustness by redundant detection across views. Object tracking, face detection, gesture analysis, etc. exploit correspondences in the available views by checking the consistency of the analysed primitives (colour, salient points...) in the various projections of the actual 3D scene.

An implicit or explicit auxiliary 3D representation in the form of a volumetric model of the scene is often used as a reference for inter-camera registration in multi-view analysis when camera calibration is available. One usually resorts to an ordered scanning of the 3D space (Cheung 2000, Kutulakos 2000), where volumetric units (or voxels) are equally sized cubes sequentially analysed from their projections in the multiple cameras.

At high resolutions, with the working 3D space sampled at regularly spaced intervals in its orthogonal axes, regular voxelization (Kaufman 1993) is adequate for volume visualization, modelling and rendering of synthetic scenes. Voxelization is also the natural support for data from direct 3D measurements in medical imaging (CT, MRI, ultrasound), biology, geosciences, industry, etc. *However, regular voxelization has a number of drawbacks for multi-view scene analysis.* This is mainly due to the fact that measurements on 3D voxels are not directly available in multi-camera settings. Voxel features are computed rather as a result of the analysis from their projections in multiple views; i.e. the analysis takes place on ‘projected’ or ‘image’ data. The actual measurements available are the data sets of pixels belonging to the voxel projections in each view.

The problem arises from the fact that the sampling geometry generated by the regular scanning of the 3D

space is distorted by the camera projection. Once projected onto the camera images, the sampling geometry becomes irregular, and the amount of data (pixels) from each view available for the analysis of each volumetric unit (voxel) depends on its distance to the camera and on the intrinsic camera parameters. Furthermore, voxel sizes (3D sampling parameters) are not dependent of image resolution and have to be carefully chosen considering the worst case (e.g. projections of two adjacent voxels should not overlap on the same pixel in most of the views). A better approach is to oversample the voxel array so that it can be guaranteed that each 3D sample is drawn from at least a single pixel (Broadhurst 2001).

Figure 1 illustrates this problem. The projections (splats) of one voxel in two different views have varying sizes for cameras located at different distances (this is the usual case for most voxels in the analysed scene). For those views where the splat size is reduced to a few pixels, image data will hardly contribute significant information to the voxel being analysed. Symmetrically, two equally sized voxels project in a different number of pixels on the same camera if they are located at different distances/orientations in 3D space.

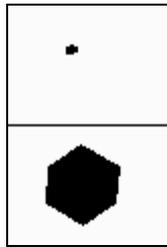


Figure 1 : Two projections of the same voxel, as seen by a close camera (top) and by a far one (bottom)

One way to overcome the dispersion in the image data for voxel analysis is to assign varying weights to the different views when analysing each voxel (Broadhurst 2001). This might result in a certain lack of ‘balance’ in data sets representing each elementary 3D unit across multiple views. Alternative approaches introduce space discretization which does not rely on regular voxels (Erol 2005) or use hybrid techniques combining volumetric and surface-based approaches (Boyer 2003).

In this paper we follow such alternative approaches and change to an irregular scanning strategy to construct the auxiliary 3D model of the scene under analysis. The resulting geometry is based on the epipolar constraint (Zhang 1998) and is proposed with the aim to better adapt 3D scene analysis to the available image data. It provides a better ‘balance’ for the analysis of volumetric elementary units from projected data. In addition, the new geometry naturally derives 3D sampling parameters from the original resolution of image data.

The following section analyses regular voxelization and provides statistical values showing the dispersion in the data used to analyse each voxel. Sections 3 and 4 define the proposed scanning geometry and outline a recursive algorithm to scan 3D space. Section 5 analyses statistics of the new geometry and Section 6 compares results obtained for an analysis technique to regular voxelization. Finally, advantages of the proposed method are discussed along with conclusions and future work.

2 Statistics of Voxel Projection Size for Regular Voxelization in Multi-view Analysis

We have analysed the problem of the dispersion in the available image data used to represent each voxel in a particular, albeit common, multi-view analysis situation. In our experiments, a Smart Room is equipped with 5 fully calibrated cameras. Four cameras are placed on the room corners and the fifth one is mounted on the ceiling, providing a zenithal view of the scene.

A regular sampling geometry with 3 cm sided cubic voxels is defined in the 3D working space of $4 \times 5 \times 2 \text{ m}^3$. We have computed the statistics of the projection size (in pixels) for all voxels in the working space. The histogram of the voxel projection size is shown in Figure 2 for one of the corner cameras (cam1). Table 1 outlines minimum, maximum, mean and dispersion values of the voxel projection size for all cameras.

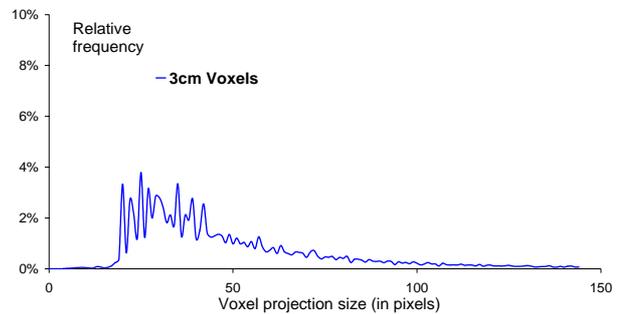


Figure 2 : Histogram of voxel projection size onto camera 1 for all $3 \times 3 \times 3 \text{ cm}^3$ voxels in a $4 \times 5 \times 2 \text{ m}^3$ working space

	<i>Min projection size</i>	<i>Max projection size</i>	<i>Mean projection size</i>	<i>Standard deviation in proj. size</i>
Cam1	4	1797	60	67
Cam2	4	1864	58	65
Cam3	4	1650	59	66
Cam4	4	2155	60	69
Cam5	2	296	40	22

Table 1 : Statistics of voxel projection size (in pixels) for 3 cm sided voxels in a $4 \times 5 \times 2 \text{ m}^3$ working space

The statistics of the voxel projection size show considerable dispersion: standard deviation is of the same order as the mean value. This situation is not favourable for multi-view analysis algorithms when looking for matching visual features in different views or when checking the consistency of the analysis across multiple projections of the actual 3D scene. The analysis algorithm will be using quite different amounts of image data (number of pixels projected in each view) for the analysis of an individual voxel in 3D space. This is due to the dispersion in the projection size of the uniform elementary unit employed in this geometry.

Regular voxelization of 3D space is, therefore, “**non-adapted**” to image data for multi-view analysis. At least in settings as common as a smart room with 5 evenly distributed cameras, voxels project to ‘unbalanced’ sets of image data in each camera, from which analysis algorithms have to work out feature matches and consistency checks. As mentioned before, one way of adapting the analysis to the available image data would

be to avoid giving the same importance to the data set (projection) in each view. When the analysis algorithm has to make a decision on that voxel (e.g. whether it is foreground or background; surface or interior; skin, clothing or object...) it will have to take into account the amount of pixels in each view informing the analysis decision. This strategy depends on the analysis itself and does not solve the lack of ‘balance’ of the data sets representing each voxel in the different views. The dispersion in the projection data sets is due to an arbitrary 3D scanning geometry chosen to support analysis data.

3 Proposed 3D Scanning Geometry

An alternative strategy is ‘image-based’ scanning of the 3D scene. Matusik introduced the concept of ‘image-based’ visual hulls (Matusik 2000, Matusik 2001) to render an observed scene in real time from a virtual camera’s point of view without constructing an explicit auxiliary volumetric representation. He claims that the advantage of performing geometric computations based on epipolar geometry in image space is the elimination of resampling and quantization artefacts in volumetric approaches. However, his paper focuses on visualization and rendering applications rather than visual analysis and does not consider the effects of image sampling. We follow Matusik’s concept of ‘image-based’ processing, but focussing on analysis applications. In particular, we propose an image-based recursive scanning algorithm for multi-view analysis, and derive the corresponding geometry in 3D space. This provides a volumetric representation for image data functionally equivalent to regular voxelization as volumetric data support for the analysis algorithms. The 3D scanning procedure is better **adapted** to the image data than regular voxelization, minimizing the dispersion in the amount of data used for the analysis of each voxel in the different views.

The motivation behind the proposed approach is that it does not make much sense to scan the 3D space (from which we do not have direct measurements) with a regular geometry while this results in a non-regular geometry once projected on the camera images. The actual data we have available in multi-view scene analysis applications are visual measurements (pixel data) from the camera images, and we better base the scene analysis geometry on the available data unless there is a clear benefit from not doing so. The proposed procedure changes the usual multi-view analysis paradigm adapting the analysis strategy to the available image data. Instead of scanning 3D space with arbitrary regular voxels –from which we do not have direct data–, the proposed scanning is natural and regular on the camera images, which are divided in a regular way, and the 3D equivalents of such divisions generate the volumetric geometry.

In the next subsections we introduce the basic tools defining a 3D geometry adapted to the images. First, we define the *quadrant* as an image region. Then, the *cone* is obtained as back-projection of the quadrant. The *conexel* –elementary volumetric unit for the proposed geometry– is obtained by intersection of cones. Finally, we outline a recursive algorithm for 3D space scanning in multi-camera settings based on this geometry, which has proven useful for multi-view analysis techniques.

3.1 Image regions: *quadrants*

To avoid dispersion in the amount of image data from the different views used in the analysis of a volumetric unit, we divide camera images in quadrants. *Quadrants* are defined as regular square shaped, non-overlapping regions in the projected images. 3D space scanning will be defined based in the geometry generated by the quadrants, instead of using the voxel-based geometry.

The expected behaviour of the proposed approach is that the data sets in every image will be balanced when scanning a 3D space region: their projections will always lie inside the selected quadrants. Furthermore, the subdivision of the images in quadrants can be made recursive, and the scanning algorithm described at the end of this section exploits this possibility.

3.2 Back-projection of quadrants in 3D: *cones*

The *cone* is the 3D back-projection of a 2D quadrant, also known as the projective extrusion of the 2D silhouette (Matusik 2000). To obtain the 3D back-projection of a quadrant in an image, we compute the back-projected ray of the four corners of the quadrant (Garcia 2005). Then, we compute the inequations of four planes by combining the four ray equations, so that the pixels in the quadrant are the projection of the inner volume enclosed by the four planes. The *Center Of Projection* (COP) of the camera is the main vertex of the cone, which results in a pyramidal shape without a basis.

An illustration of two such cones computed from their corresponding quadrants is shown in Figure 3 for the actual settings of cameras 1 and 2 in our smart room.

3.3 Intersection of two or more cones: *conexels*

The elementary volumetric unit in our scanning geometry is called *conexel*¹. We define the *conexel* as the 3D intersection of back projected cones. The cones defining a *conexel* are generated by a selected quadrant in each available view. The procedure to obtain a *conexel* is:

1. Select a quadrant for every available camera image
2. Compute back-projected cones for the selected quadrants (a set of 4 inequations define each cone)
3. Obtain volumetric intersection of computed cones

Figure 4 presents a 3D view of a *conexel* obtained as the intersection of the two cones shown in Figure 3, corresponding to quadrants (2,1) and (1,1) selected from the views in cameras 1 and 2, respectively. Clearly, the geometry of the *conexel* is that of a polyhedral visual hull and its 3D computation is perhaps not so straightforward. We will see that the defined geometry will be implicitly used in the proposed scanning algorithm and, unless an explicit volumetric representation with *conexels* is required, multi-view scene analysis algorithms do not need to compute the 3D *conexels*. Anyway, computing and rendering a 2D ‘view-dependent’ representation of a polyhedral hull can be done efficiently (Matusik 2001).

¹ Named after ‘cone element’ in analogy with pixel from ‘picture element’ and ‘voxel’ from ‘volume element’

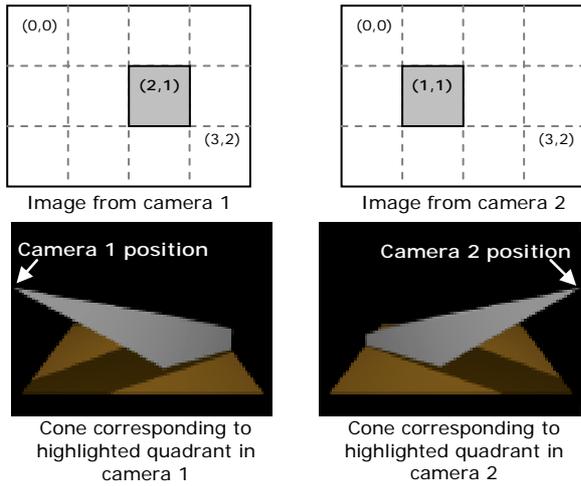


Figure 3 : Two *cones* (bottom) computed as projective extrusions of the two camera *quadrants* (highlighted at top). Room floor (colored square) included as visual reference



Figure 4 : *Conexel* (bottom) obtained as the intersection of two *cones* (top). Room floor is again included as reference

4 Scanning Method

For multi-view scene analysis purposes, the projection of a *conexel* in every camera can be computed in a fast way, without having to calculate its actual 3D geometry and project it on every camera image. This is accomplished using epipolar geometry.

In particular, we compute equations of the epipolar lines corresponding to the corners of each quadrant in the other views using fundamental matrices (Hartley 2000). The equations of the epipolar lines are converted to inequations (Ma 2003) so that we can define two image regions in the current view for every cone generated by the quadrant in another view: pixels lying inside the cone projection and pixels lying outside. The area limited by the inequations generated by all epipolar lines defines the projection of the intersection of the cones generated by the quadrants from the other views in the current view (see Figure 5). Pixels inside the quadrant in the current view for the working camera are checked and only those also lying inside the projections of all cones are selected as belonging to the projection of the *conexel* on the current camera image.

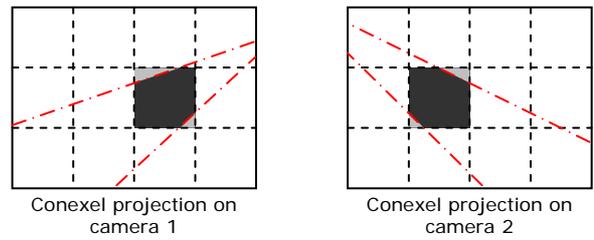


Figure 5 : In general, the reprojection of a *conexel* (dark grey) does not completely cover the generating *quadrants* (light grey). The innermost epipolar lines (corresponding to the corners of the quadrant in the other view) are shown in red for each view

Please note that, when projecting the *conexel* obtained as the intersection of cones onto the camera images, the projections do not completely cover the quadrants which generated the *conexel*, as shown in Figure 5. As a consequence, when using the proposed geometry with the *conexel* as elementary scanning volume, there will still be some dispersion in the amount of image data used for the analysis of the volumetric unit in 3D space. Anyway, the dispersion is expected to be smaller than with regular voxelization. In fact, the number of pixels of the projection of the *conexel* in each camera view will range from 1 to the total number of pixels in the quadrant, with a dispersion range usually much smaller than the dispersion range for regular voxelization computed in section 2. We will present statistics to assess this statement in a quantitative manner in the results section, proving that the proposed image-based 3D geometry is better adapted to the image data, and provides a better balance in the sets of image data (pixels) characterizing the volumetric elementary unit across the available views.

4.1 Recursive scanning and the *m-tree*

The proposed scanning method based on quadrants can be implemented in a recursive 3D space scanning algorithm, allowing progressive scene analysis approaches. By performing a quad-tree decomposition on the projected 2D image data, each quadrant can be subdivided in four sub-quadrants. The algorithm proceeds by dividing the resulting quadrant in sub-quadrants until some analysis condition is met (e.g. until a foreground or colour consistency check yields true). For each division, the result will be a new set of *conexels*, always included in the previous one. This strategy can be used to selectively enhance the resolution of 3D analysis only in the regions where needed –such as objects contours– without using the highest resolution in homogeneous space regions, where it is not necessary to subdivide further. Therefore, a progressive space analysis algorithm based on the proposed procedure may start the analysis at rough resolution levels using large quadrants (*conexels*) and progressively refine the analysis by recursive subdivision to scan at higher resolution only those quadrants (*conexels*) where needed, depending on the analysis results at the previous resolution level.

Figure 6 illustrates the recursive subdivision and its representation in an *m-tree*. The *m-tree* (Lu 1996) has been chosen to store the progressive analysis results of the recursive scanning algorithm. Its implementation includes a set of functions which allow moving up, down and sideways in the tree structure.

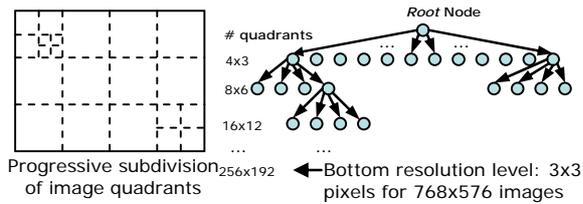


Figure 6 : Progressive scanning of 3D space by recursive subdivision of quadrants and its representation in an m-tree

4.2 Algorithm allowing 3D recursive scanning for progressive multi-view analysis

As the number of cameras may vary, the proposed algorithm will loop depending on the number of cameras. A vector is defined to store the current quadrant under analysis for every camera and a variable stores the current camera in use. The main recursive algorithm to scan 3D space based on the geometry defined by the conexels works through the following steps:

1. Set the current camera view to zero (first camera) and the chosen quadrant vector to all zeros.
2. If the currently selected camera index is larger or equal to zero go to the next step; otherwise, finish.
3. For each camera with smaller index than the one currently selected, select one quadrant and obtain their cone projections onto the currently selected camera view. Also obtain cone projections for the current quadrant in the currently selected camera onto the camera views with smaller index. If any of the cameras does not have any pixel belonging to the conoxel projection, it means that no conoxel exists for the current set of quadrants. In that case jump to step 6; otherwise, go to the next step.
4. If the currently selected camera is the last one available (meaning that a conoxel exists for the current set of quadrants), count the number of pixels of the conoxel projection on every camera and, if different from zero, go to next step. In any other case, select next camera and jump to step 2.
5. At this step **any visual analysis function** can be implemented requiring a multi-view consistency check on the projected pixels corresponding to the obtained conoxel in 3D. In case that the consistency check needs a higher resolution, jump to step 7; otherwise, store the results in the *m-tree* and go to the next step.
6. If the current quadrant in the current camera is not the last one, increment it. Otherwise, set it to 0, decrement the currently selected camera index and repeat this step while the current camera is larger or equal than zero. Finally jump to step 2.
7. Subdivide each quadrant in smaller quadrants in every view, go down in the *m-tree*, call recursively this procedure and go up in the *m-tree* again². Then jump to step 6.

² The available quadrants in every camera are stored in the *m-tree*

5 Statistics of Conoxel Projection Sizes for the Proposed Scanning Geometry

As stated before, an improvement of analysis results is expected due to the fact that the presented scanning approach is more natural to image data. In particular, the proposed geometry minimizes the dispersion in the number of pixels used for the analysis of the elementary volumes when projected onto the different camera views.

As the projections of the conoxel onto the camera images do not completely cover the quadrants, the projection size of the elementary volumetric unit of the proposed geometry is not constant. We have compared the distribution of the conoxel projection size in the same 3D working space with those of regular voxelization shown in section 2. In order to compare the statistics of the two geometries, we note that the average projection size for 3 cm sided voxels is 60 pixels for camera 1 (see Table 1). This value is in between the projection sizes of 6x6 pixels and 12x12 pixels quadrants. This is why we show the distribution for these two cases in Figure 7.

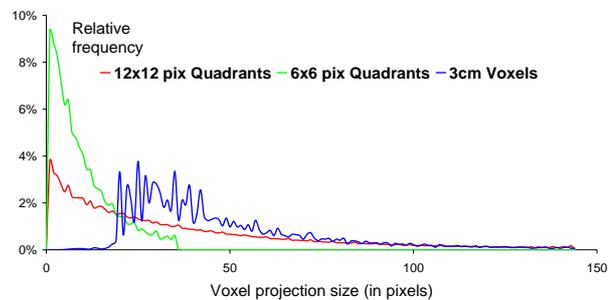


Figure 7 : Histogram of voxel projection size for 3 cm sided voxels (blue line, same as Figure 2), quadrants of 6x6 pixels (green) and quadrants of 12x12 pixels (red)

The distribution of the conoxel projection size has completely changed its shape compared with regular voxelization. The range of possible values is restricted to the maximum quadrant size³ and dispersion values are reduced with respect to the regular case, but standard deviation is still of the order of the mean value of the distribution.

6 Experimental Results

In this section we provide an objective validation proving that the proposed geometry is better adapted to image data in terms of sampling accuracy. This will serve as a proof of concept aiming to quantitatively evaluate the extent to which the geometry based on conexels improves analysis applications in multi-camera settings. Then, we illustrate the progressive capabilities of the proposed multi-view scanning algorithm in a real application.

The analysis application chosen for the experiments is 3D foreground segmentation or Shape-from-Silhouette (Landabaso 2005), which has been designed for 3D object tracking in the smart room.

³ Note: For these measures, we assume that the recursive algorithm goes always down to the highest resolution (either 6x6 or 12x12 pixels) for all quadrants in all views.

In Shape from Silhouette applications input data (camera views) are binary images obtained as foreground segmentation masks by a 2D foreground extraction algorithm (Stauffer 2000) from the original camera views. In experiments with real image data, inaccuracies of 2D foreground extraction⁴ might prevent an exact quantitative evaluation of the performance of the proposed geometry. This is why we have first chosen the projections of an ideal object (a sphere) in order to have the ground truth available for quantitative comparison.

6.1 Proof of concept: synthetic sphere

For this proof of concept, we have generated 5 simple synthetic scenes. A sphere with a diameter of 1 meter is placed at 5 different positions in the working space of the smart room. As ground truth, we generate the images for the camera views by projecting with the actual intrinsic and extrinsic parameters of every camera. An example of one of these input projections is illustrated in Figure 8.

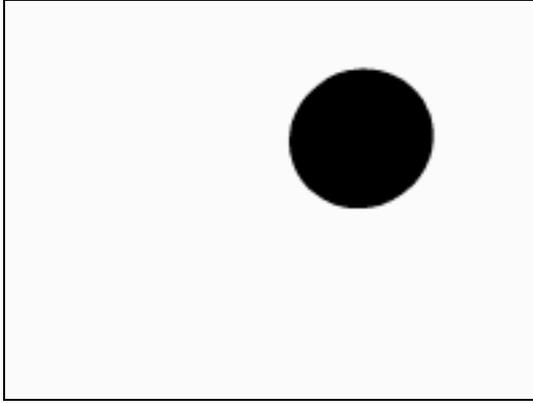


Figure 8 : Input projection for camera 2 for the first of the 5 generated scenes with a simple synthesized sphere. We will take this as ground truth because, contrary to real foreground scenes, it does not show noise effects, misses or false detections

We compare the presented image-based scanning approach for 3D foreground segmentation algorithm with regular voxelization as the competing method. As before, the criteria for comparing results is using a voxel side that, in average, has a projection size in pixels on camera images equal to the number of pixels of the smallest quadrant used in the reconstruction. The formula

$$\begin{aligned} nPix &\approx \pi r^2 \\ &\approx \pi \left(\frac{\sqrt{3}/2 \cdot voxSide \cdot K_0}{dist(voxCenter, COP)} \right)^2 \\ &= \alpha \cdot voxSide^2 \end{aligned}$$

allows computing an equivalent voxel side from the number of pixels we want the voxel to be projected to. We approximate the average voxel projection size in pixels by the projection size of the central voxel in the sphere, and compute the α in the formula for all the scenes and for every camera. Then, an average of α is computed for every scene along all the available cameras.

⁴ like misses or false detections, and the presence of noise appearing as isolated or grouped foreground pixels

Synthetic sphere scene num. (average α)	Equivalent voxel side for quadrants of		
	3x3 pixels	6x6 pixels	12x12 pixels
Scene 1 ($\alpha=8.256$)	1.04 cm	2.088 cm	4.176 cm
Scene 2 ($\alpha=8.222$)	1.05 cm	2.092 cm	4.185 cm
Scene 3 ($\alpha=8.160$)	1.05 cm	2.100 cm	4.201 cm
Scene 4 ($\alpha=8.226$)	1.05 cm	2.100 cm	4.184 cm
Scene 5 ($\alpha=7.104$)	1.03 cm	2.251 cm	4.502 cm
Voxel side size taken:	1 cm	2 cm	4 cm

Table 2 : Equivalent voxel size side for various quadrant sizes

For quadrants of 3x3 pixels, the equivalent voxel side in all scenes is around 1.1 cm. So, to be in the safe side, we take 1 cm as the equivalent voxel side. This is done similarly for quadrants of 6x6 and 12x12 pixels, as shown in Table 2, resulting in equivalent voxel sides of 2 cm and 3 cm respectively, with a slight advantage in resolution for regular voxelization in all cases. After performing the analysis with both methods, the 3D foreground volume reconstructed is projected back to all cameras.

To evaluate the distortion in the projected image introduced by the sampling 3D geometry with respect to the original ground truth, we define the following metric:

$$dist(rec, gt) = \frac{area(rec \cup gt) - area(rec \cap gt)}{area(gt)}$$

that is, the distance is computed as number pixel differences among reconstructed projection and ground-truth divided by the number of pixels of ground-truth. This distance function is computed for every available projection of the sphere.

In the case of 3x3 pixel quadrants, Table 3 and Table 4 list the distance to ground truth for all 5 scenes and 5 cameras for regular voxelization and image-adapted scanning. Results are given in %, with a multiplicative factor of 100 to render them more easily readable.

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
Cam1	7.40	7.28	7.65	7.57	7.46	7.47
Cam2	7.25	7.46	7.62	7.74	7.34	7.48
Cam3	7.68	7.63	7.46	7.24	7.52	7.51
Cam4	7.62	7.65	7.39	7.49	7.35	7.50
Cam5	7.31	7.46	7.35	7.38	7.47	7.39
Average	7.45	7.50	7.49	7.48	7.43	7.47

Table 3 : Distance to ground-truth for regular voxelization with 1 cm sided voxels

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
Cam1	2.03	2.55	3.12	2.91	2.56	2.63
Cam2	2.51	1.95	3.13	2.98	2.83	2.68
Cam3	3.29	2.97	1.89	2.37	2.76	2.66
Cam4	2.74	3.19	2.50	2.01	2.81	2.65
Cam5	2.36	2.37	2.66	2.74	2.52	2.53
Average	2.59	2.61	2.66	2.60	2.70	2.63

Table 4 : Distance to ground-truth for image-adapted scanning with 3x3 pixel quadrants

Image-adapted scanning provides more accurate reconstruction. The averaged pixel differences (or errors) are about 35% of those obtained for regular voxelization with quadrants of size 3x3 pixels.

To illustrate those results, Figure 9 shows pixel differences between the ground-truth image and the projection from the reconstructed 3D object for the third scene projected on camera 2 obtained with regular voxelization. Figure 10 is the equivalent result with image-adapted scanning with conexels. Please note how the dispersion in the amount of data used for analysis in regular voxelization causes larger false volumes to appear.

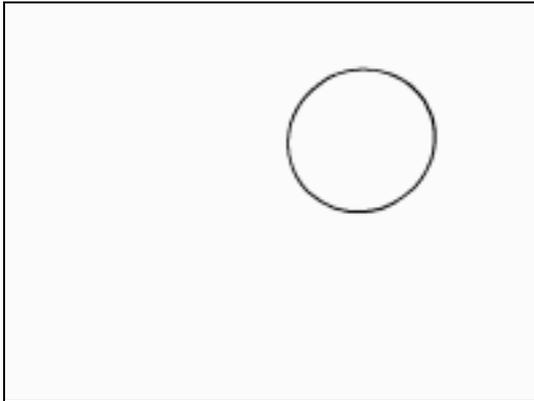


Figure 9 : Pixel differences between the reconstruction with regular voxelization and ground-truth for the third scene on camera 2 (voxel side 1 cm)



Figure 10: Pixel differences between the reconstruction with image-adapted scanning and ground-truth for the third scene on camera 2 (3x3 pixels quadrants)

The cases of 6x6 and 12x12 pixel quadrants yield similar results. We have just provided the averaged metrics per scene for two pairs of equivalent cases in Table 5 and Table 6 for regular voxelization and image-adapted scanning. Please observe that, at these lower resolutions, image-adapted scanning still provides more accurate reconstruction. Pixel differences are now about 50% of regular voxelization. Figure 11 through Figure 14 show pixel differences between ground-truth and projection of the reconstructed sphere for regular and image-adapted scanning for the cases of 2 cm and 4 cm sided voxels, and the equivalent cases of 6x6 pixels and 12x12 pixels quadrants.

<i>All cameras</i>	<i>Scene 1</i>	<i>Scene 2</i>	<i>Scene 3</i>	<i>Scene 4</i>	<i>Scene 5</i>	<i>Average</i>
Regular with 2 cm sided voxels	13.23	13.13	13.12	13.11	13.03	13.12
Image adapted with 6x6 pixels quadrants	6.31	6.68	6.31	6.47	6.86	6.53

Table 5 : Distance to ground-truth for regular voxelization with 2 cm sided voxels and image-adapted scanning with (equivalent) 6x6 pixels quadrants

<i>All cameras</i>	<i>Scene 1</i>	<i>Scene 2</i>	<i>Scene 3</i>	<i>Scene 4</i>	<i>Scene 5</i>	<i>Average</i>
Regular with 4 cm sided voxels	24.21	24.22	24.12	24.18	24.03	24.15
Image adapted with 12x12 pixels quadrants	14.34	14.43	13.17	13.98	14.66	14.11

Table 6 : Distance to ground-truth for regular voxelization with 4 cm sided voxels and image-adapted scanning with (equivalent) 12x12 pixels quadrants

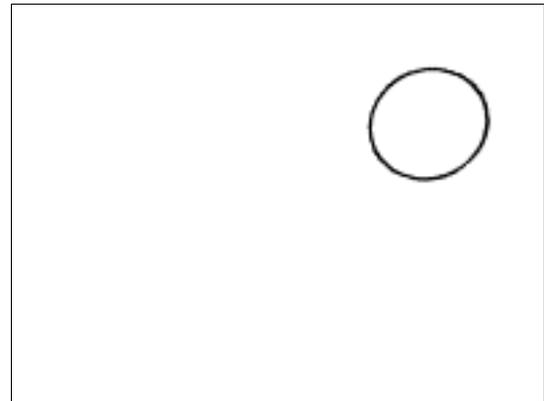


Figure 11: Pixel differences for the reconstruction with regular voxelization (voxel side 2 cm)

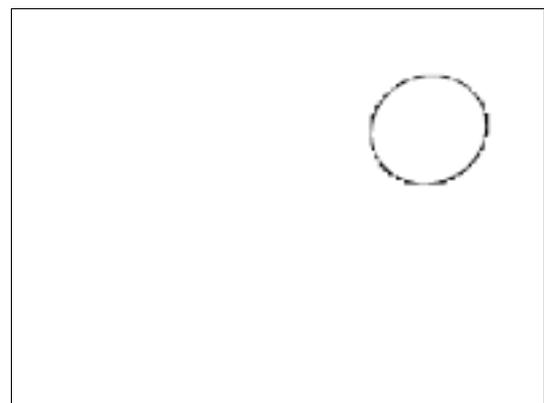


Figure 12: Pixel differences for the reconstruction with image-adapted scanning (6x6 pixels quadrants)

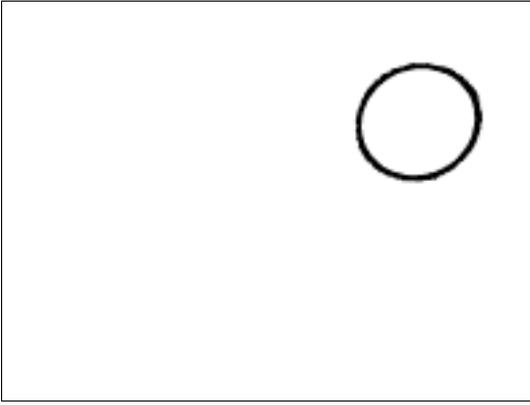


Figure 13: Pixel differences for the reconstruction with regular voxelization (voxel side 4 cm)



Figure 14: Pixel differences for the reconstruction with image-adapted scanning (12x12 pixel quadrants)

6.2 Real SfS application

The proof of concept provided above for the reconstruction of a synthetic sphere from its silhouette projections has shown quantitative improvements for the image-based scanning method. We now show qualitative results for the proposed 3D scanning geometry with actual images from our smart room, in a real application of the Shape-from-Silhouette multi-view analysis algorithm. In addition, we illustrate the progressive performance of the multi-view scanning algorithm introduced in section 4.2 in 3D foreground segmentation for object tracking (Landabaso 2005).

Figure 15 and Figure 16 show the re-projected masks obtained as result of ‘progressive’ Shape-from-silhouette reconstruction by checking foreground consistency in 5 cameras with increasing resolutions in quadrants sizes. Please note that we start at the lowest resolution with only $4 \times 3 = 12$ quadrants of 192×192 pixels each. We apply the consistency check in all views for each conoxel, as proposed in the Shape-from-Silhouette technique (Landabaso 2005), to see whether the given conoxel is all background, all foreground or mixed. Only in the later case when the conoxel is partly foreground and partly background, we continue the recursion at lower resolution by subdividing the quadrant to the next resolution step. As soon as a conoxel is detected as uniform (either all foreground or all background), the progressive analysis stops. In these settings, most conoxels are only

background and this efficiently saves further consistency analysis for such ‘uniform’ conoxels.

Figure 17 shows the results of a different progression of the 3D scanning algorithm also for Shape-from-silhouette reconstruction. In this case, we increase the number of cameras, starting from an initial reconstruction with 2 cameras and then adding the rest one by one. Of course, the two dimensions of progressive analysis (increasing resolution, increasing number of cameras) can be combined at will. The *m-tree* data structure has proven to be a valuable tool to store the data in progressive analysis strategies.

7 Conclusions and Future Work

We have presented an image-based multi-view analysis approach using a 3D space scanning geometry which is adapted to the images. Instead of exploring 3D space (from which we do not have direct measurements, but only projections) with regular geometry, the proposed scanning procedure defines a geometry based on image quadrants. The geometry builds on the concepts of image quadrant, its volumetric extrusion (the cone) and the intersection of two cones (the conoxel). This strategy adapts the multi-view analysis to the available data (pixels in camera images), improving the accuracy of the analysis from the multiple views. Contrary to the arbitrary choice of a voxel size in regular voxelization, the sampling geometry in 3D is naturally derived from the resolution of the camera images. Furthermore, volumetric scanning can be progressively refined as the analysis proceeds.

The results obtained show less dispersion in the data sets from the multiple views used to inform analysis decisions for each elementary volumetric unit. As a drawback, we must remark that dispersion in the amount of data used in analysis has not been completely cancelled, but it is more controlled than with regular voxelization techniques. The results also show increased spatial accuracy when compared with regular voxelization. This is the expected behaviour because of the balanced usage of the directly measured data. With the proposed geometry, we do not have to select a voxel size for the working space depending on the smallest splat in the projection of the elementary voxels. The size of the elementary volumetric unit is a consequence of the analysis of image-data (the smallest quadrant).

Furthermore, a recursive algorithm based on the proposed 3D scanning geometry has been described. An interesting feature of the proposed algorithm is its capability for progressive analysis, either by adaptively increasing spatial resolution (subdividing quadrants from larger to smaller sizes when needed), or by adding new cameras to the analysis as their views are made available.

The main directions for future improvements must focus in the study of the connectivity of neighbouring conoxels, and in how to use connectivity to remove inner conoxels from analysis results in case of need. Another line of study is the set of situations in which conoxels are defined from a smaller number of cameras to deal with cases where a conoxel is only visible in a subset of all the available cameras.

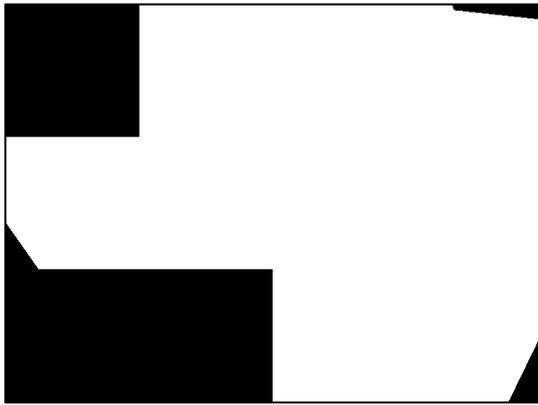


Figure 15: Results of progressive Shape-from-silhouette reconstruction by checking foreground consistency in 5 cameras with increasing resolutions in quadrants sizes. From top to bottom: 192x192, 96x96, 48x48 and 24x24 pixels per quadrant (continued in Figure 16)

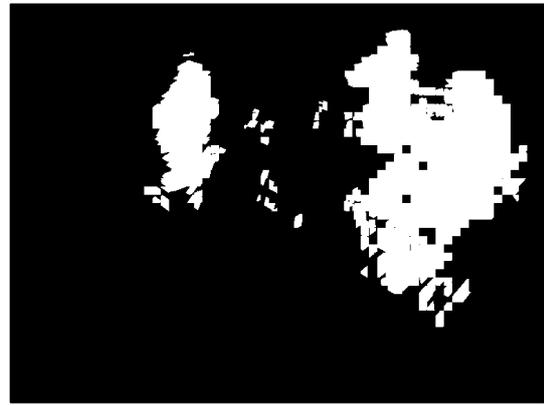


Figure 16: (continued from Figure 15) Results of progressive Shape-from-silhouette reconstruction by checking foreground consistency in 5 cameras with increasing resolutions in quadrants sizes. From top to bottom: 12x12, 6x6 and 3x3 pixels per quadrant. The last image is the original image with its original lens distortion (corrected in a pre-processing step).



Figure 17: Results of progressive Shape-from-silhouette reconstruction with increasing number of cameras. From top to bottom: projection of the 3D reconstruction with 2, 3, 4 and 5 cameras. Bottom image: original (noisy) 2D foreground.

References

- Boyer, E., Franco, J.-S. (2003) 'A hybrid approach for computing visual hulls of complex objects', in *Computer Vision and Pattern Recognition (CVPR'03)*, vol.1, pp. 695–701.
- Broadhurst, A., Drummond, T.W., Cipolla, R. (2001), 'A Probabilistic Framework for Space Carving', in *Proc. 8th International Conference on Computer Vision (ICCV'01)*, vol.1, pp. 388–393.
- Cheung, G., Kanade, T., Bouguet, J.-Y. & Holler, M. (2000), 'A real time system for robust 3d voxel reconstruction of human motions', in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, Vol. 2, Hilton Head Island, South Carolina (US), , pp. 714–720.
- Erol, A., Bebis, G., Boyle, R. D., Nicolescu, M. (2005), 'Visual Hull Construction Using Adaptive Sampling', in *7th IEEE Workshops on Application of Computer Vision (WACV'05)*, vol.1.
- Garcia, O. & Casas, J. (2005), 'Functionalities for mapping 2D images and 3D world objects in a multi-camera environment', in *Proc. 6th Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05)*, Montreux, Switzerland, pp. 241–249.
- Hartley, R. & Zisserman, A. (2000), *Multiple view geometry in computer vision*, Cambridge Univ.Press.
- Kaufman, A., Cohen, D., & Yagel, R. (1993) 'Volume Graphics'. *IEEE Computer* **26** (7), pp. 51–64.
- Kutulakos, K.N., Seitz, S.M (2000), 'A Theory of Shape by Space Carving', *International Journal of Computer Vision* **38** (3), pp. 199–218.
- Landabaso, J. & Pardas, M. (2005), 'Foreground regions extraction and characterization towards real-time object tracking', in *Proc. Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'05)*, Edinburgh, UK.
- Lu, T. (1996), 'The enumeration of trees with and without given limbs', *Discrete Math*, 154 (1-3), pp. 153–165.
- Ma, Y., Soatto, S., Kosecka, J. & Shankar-Sastry, S. (2003), *An Invitation to 3-D Vision: From Images to Geometric Models*, Springer Verlag.
- Matusik, W., Buehler, C., McMillan, L. (2001), 'Polyhedral visual hulls for real-time rendering', in *Proc. 12th Eurographics Workshop on Rendering EGWR'01*, London.
- Matusik, W., Buehler, C., Raskar, R., Gortler, S.-J. & McMillan, L. (2000), 'Image-based visual hulls', in *Proc. 27th conf. on Computer graphics & interactive techniques (SIGGRAPH'00)*, pp. 369–374.
- Puech, W., Bors, A.G., Pitas, I., Chassery, J.-M. (2001) 'Projection Distortion Analysis for Flattened Image Mosaicing from Straight Uniform Generalized Cylinders', *Pattern Recognition* **34** (8), pp. 1657–1670.
- Stauffer, C. & Grimson, W. (2000), 'Learning patterns of activity using real-time tracking', *IEEE Trans. on PAMI* **22** (8), pp. 747–757.
- Zhang, Z. (1998), 'Determining the Epipolar Geometry and its Uncertainty: A Review', *International Journal of Computer Vision* **27** (2), pp. 161–198.

Image Feature Evaluation for Contents-based Image Retrieval

Adam Kuffner¹ and Antonio Robles-Kelly^{2,3}

¹ Department of Theoretical Physics, Australian National University, Canberra, Australia

² Vision Science, Technology and Applications (VISTA), NICTA*, Canberra, Australia

³ Department of Information Engineering, Australian National University, Canberra, Australia

¹u3966239@anu.edu.au ²antonio.robles-kelly@nicta.com.au

Abstract

This paper is concerned with feature evaluation for content-based image retrieval. Here we concentrate our attention on the evaluation of image features amongst three alternatives, namely the Harris corners, the maximally stable extremal regions and the scale invariant feature transform. To evaluate these image features in a content-based image retrieval setting, we have used the KD-tree algorithm. We use the KD-tree algorithm to match those features corresponding to the query image with those recovered from the images in the data set under study. With the matches at hand, we use a nearest neighbour approach to threshold the Euclidean distances between pairs of corresponding features. In this way, the retrieval is such that those features whose pairwise distances are small, “vote” for a retrieval candidate in the data-set. This voting scheme allows us to arrange the images in the data set in order of relevance and permits the recovery of measures of performance for each of the three alternatives. In our experiments, we focus in the evaluation of the effects of scaling and rotation in the retrieval performance.

1 Introduction

The contents-based retrieval in image databases is a daunting and often costly task. As in a traditional database, the image must be described in order to be incorporated to the database and form part of the index. The engine of the database uses the index to quick-search the most probable candidates and then, making use of a similarity measure, retrieves the candidate images in order of relevance.

Image retrieval has been a long standing problem in computer vision and pattern recognition and early surveys can be tracked back to the mid-eighties (Tamura & Yokoya 1984). Nonetheless, one of the first attempts to cast the problem of retrieving images from a database as a task based upon content was that introduced in (Sclaroff & Pentland 1993). Here, Sclaroff and Pentland presented a method in which the user is allowed to provide a search model, such as a sketch or example image so as to perform a query whose output is a set of thumbnails ordered by relevance. In their model, the concept of “relevance” implies similarity, which is modeled as a continuous value between zero and unity. This measure is often modeled as

*National ICT Australia is funded by the Australian Governments Backing Australia’s Ability initiative, in part through the Australian Research Council.

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

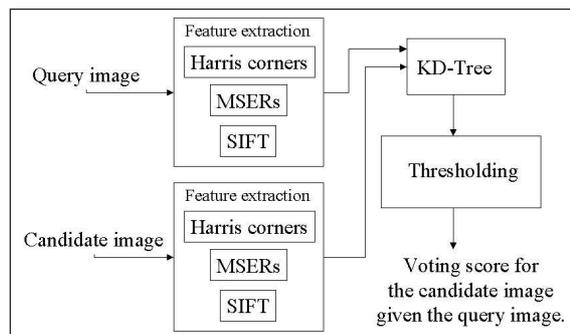


Figure 1: Diagram of our voting score recovery scheme.

a metric based on image features such as colours, corners, edges, etc.

Following this trend, automatic image database systems use elementary features in the image to index and retrieve the candidates from the database. For instance, QBIC (Niblack et al. 1993) allows images to be retrieved using shape, colour and texture. FourEyes (Picard 1995) employs a high-level image feature processing scheme to modify the structure of the database and the retrieval parameters. Photobook (Pentland, Picard & Sclaroff 1994) is a collection of tools to search and organise image datasets. SQUID (Shape Queries Using Image Databases) (Farzin & Kittler 1996) uses a scale space representation of shape so as to accomplish queries based upon contour similarity.

The main argument levelled against these systems concerns their lack of robustness to rotation and occlusion. Also, they often require a human expert to determine the parameters of the search criteria. As a result, retrieval methods vary greatly from one application to another and currently available image database systems make use of a hash or index to retrieve the images. Furthermore, the results of the query operation and the performance of retrieval applications rely heavily on an appropriate selection of the image features used to characterise the images under study.

As a result, recently, there has been an increasing interest in the evaluation of image features and descriptors computed from interest points on the image. For instance, Caneiro and Jepson (Caneiro & Jepson 2002) have used Receiver Operating Characteristics (ROC) to compare test query descriptors against a library of reference computed from a separate dataset. Mikolajczyk and Schmid (Mikolajczyk & Schmid 2005) have evaluated a number of local descriptors in the context of matching and recognition. Mikolajczyk *et al.* (Mikolajczyk, Tuytelaars, Schmid, Zisserman, Matas, Schaffalitzky, Kadir & Gool n.d.) have taken this analysis further and evaluated local descriptors subject to affine transformations. In a series of related developments, Randen and Husoy (Randen & Husoy 1999) and Varma and Zisserman (Varma & Zisserman. 2003) have compared different local image descriptors, also known as filters, for texture classification.

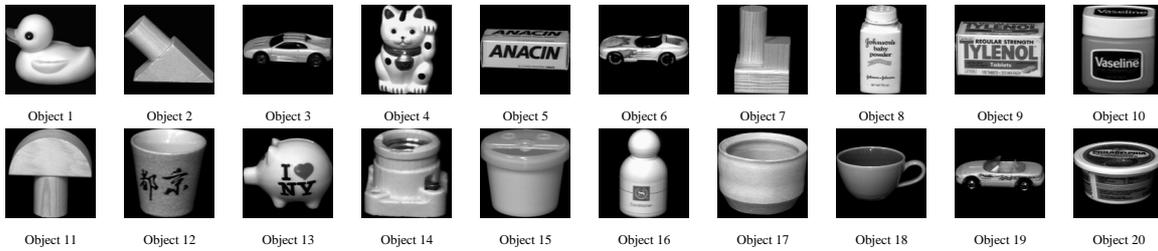


Figure 2: Sample views for the objects in the Columbia University COIL database.

2 Image Feature Evaluation and Contents-based Image Retrieval

In contrast with other studies elsewhere in the literature, here we compare the performance of local descriptors for purposes of contents-based image retrieval. Rather than evaluating the capabilities of the image features to describe the scene subject to affine transformations, we focus on the adequacy of the local descriptors for contents-based image retrieval and compare them using the same evaluation criteria and experimental vehicle. To this end, we have selected three alternatives which have previously shown a good performance in such a context and evaluate the retrieval rate under viewpoint rotation and image scaling.

As mentioned earlier, in this paper we aim to evaluate the adequacy for content-based image retrieval of three feature recovery methods. These are the Harris corner detector (Harris & Stephens 1988), the Maximally Stable Extremal Regions (MSERs) (Matas, Chum, Martin & Pajdla 2002) and the Scale Invariant Feature Transform (SIFT) (Lowe 2004). Thus, we have divided the section into two parts. The first of these concerns an overview of the local image descriptors used as alternatives for the recognition process. The second of these introduces the retrieval scheme used for purposes of the evaluation presented in this paper.

2.1 Retrieval Process

Having provided an overview of the image features to be evaluated, we now present the image retrieval scheme used throughout the paper for the purposes of evaluating the suitability of the local descriptors above for purposes of content-based image retrieval. The diagrammatic representation of our retrieval scheme is shown in Figure 1. Our method recovers, at input, the features for both, the images in the database and the query image. With the image features at hand, we recover the correspondences between features in both, the query and each of the data images making use of the KD-tree algorithm (Bentley 1975). These correspondences are an equivalence relation which we use to recover a score that depicts the similarity between the query and the data images. This score is based upon the Euclidean distances between pairs of corresponding features.

Being more formal, consider the query image I_Q and the data image I_D whose respective feature sets are $\Omega_Q = \{\omega_Q(1), \omega_Q(2), \dots, \omega_Q(|\Omega_Q|)\}$ and $\Omega_D = \{\omega_D(1), \omega_D(2), \dots, \omega_D(|\Omega_D|)\}$, where $\omega_Q(i)$ and $\omega_D(j)$ are the i^{th} feature vectors for the model and the data images. If there is a match between the feature vectors $\omega_Q(i)$ and $\omega_D(j)$, their squared Euclidean distance $r(\omega_Q(i), \omega_D(j))$ can be used to recover the score

$$\beta = \frac{|\Gamma|}{\max\{|\Omega_Q|, |\Omega_D|\}} \quad (1)$$

where Γ is the set of feature vectors recovered from the data image I_D whose pairwise distances with respect to

their matching query-image features are below a given threshold ϵ , i.e. $\omega_D(i) \in \Gamma \iff r(\omega_Q(i), \omega_D(j)) \leq \epsilon$.

As a result, if $\beta = 1$, the feature vectors recovered from the query image are all sufficiently “close” to the features in the data image. Furthermore, the number of features for the query and data images must be equal, i.e. $|\Omega_Q| = |\Omega_D|$. On the other hand, if β tends to zero, then the number of feature vectors in the query image which are far apart from those features in the data image to which they have been matched is large. Hence, β is a normalised “voting” score which can be viewed as a similarity measure between the query and the data image.

Furthermore, β captures the similarity between images based upon their features. Thus, the retrieval is based upon the image “contents”. To construct the feature vectors $\omega_Q(i)$ and $\omega_D(i)$ we have considered the nature of each of the three alternatives evaluated here. Furthermore, by construction, the feature vectors are a set of parameters that describe the feature under study. For instance, recall that the Harris Corner detector finds specific corners on objects and features within a grey scaled image. It does this by taking an image and convolving it with a Sobel gradient filter to produce gradient maps, which are then used to compute the locally averaged moment matrix. It then combines the eigenvalues of the moment matrix to compute a “corner strength”, of which maximum values indicate the corner positions. Thus, in our experiments, our feature vector corresponds to the x and y coordinates on the image plane for the detected corners.

In the case of the MSERs, the algorithm operates on a grey-scale image by finding the regions that are maximally stable with respect to changes in pixel intensities. With the MSERs at hand, we fit an ellipse to each of the recovered regions making use of the algorithm of Fitzgibbon, Pilu and Fisher (Fitzgibbon, Pilu & Fisher 1999), which fits ellipses to the recovered MSERs so as to minimise the sum of squared algebraic distances. Thus, for the MSERs, our feature vector is a five-dimensional one comprised by the centroid coordinates, orientation and major and minor axis lengths of the ellipse fitted to each of the maximally stable regions.

In contrast with the Harris corners and MSERs, where the feature vector is based upon the geometric interpretation of the aim of computation of the feature recovery method, in the case of the SIFT, we make use of the 128-element descriptor yield by the method of Lowe (Lowe 2004). The SIFT recovers and characterises points invariant to scaling making use of a four-stage cascading filter approach which commences with a scale-space extrema detection. This first step consists of a difference-of-Gaussians, which is used to identify potential interest points. Keypoint localisation is then used to eliminate previously calculated keypoints that, either have low contrast or are not localised on an edge. After recovering keypoint orientations, local gradient data is used to construct a descriptor for each of the recovered keypoints.

3 Results

As mentioned earlier, our aim here is the evaluation of the three feature descriptors above for purposes of contents-

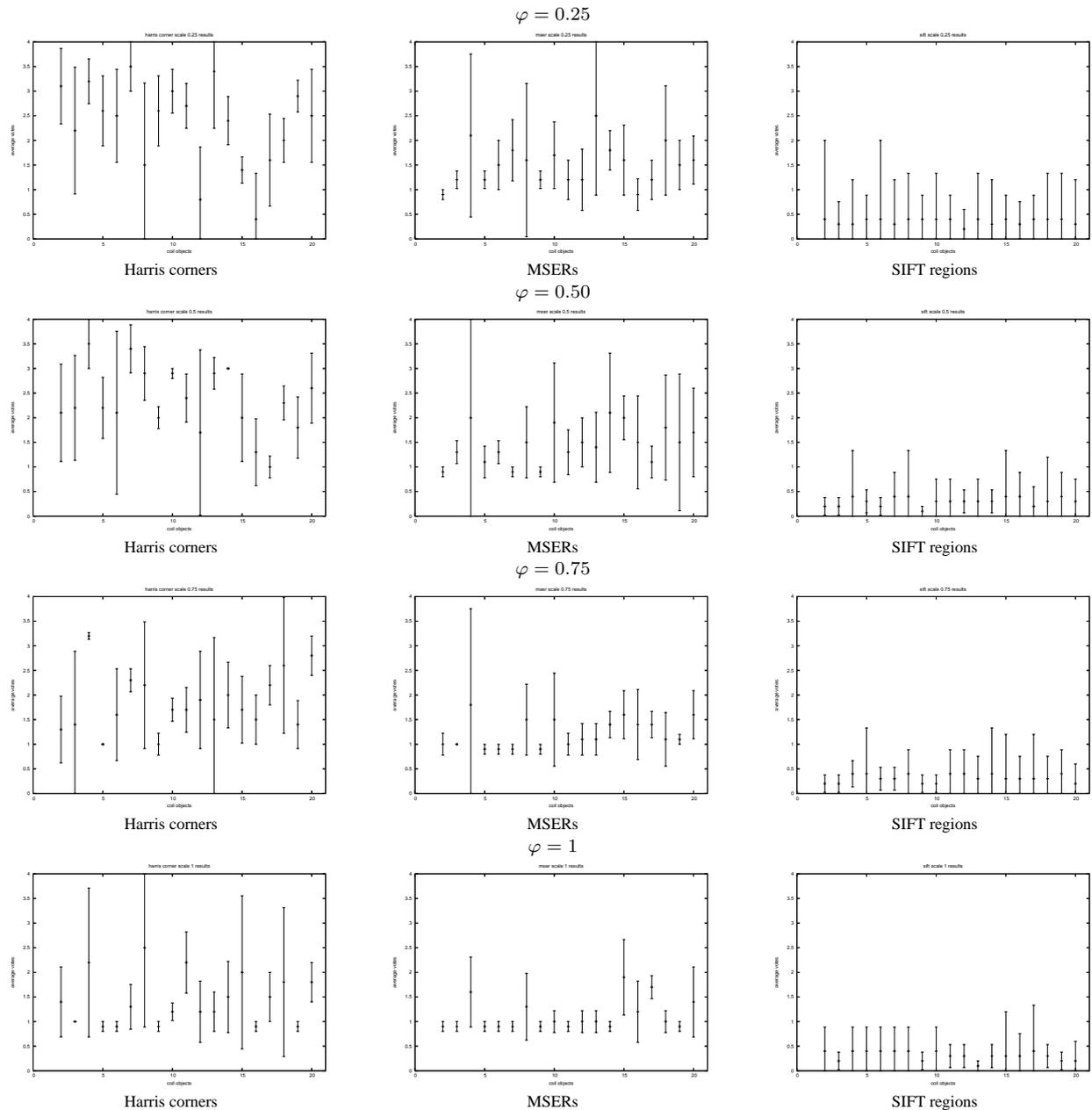


Figure 3: Retrieval results for the three alternatives of image feature vectors and four different values of the scaling factor φ .

based image retrieval. Thus, in this section, we assess the quality of the retrieval results using, as an experimental vehicle, the Columbia University COIL-20 and an in-house acquired database of urban scenes. To evaluate the effect of scaling on the image retrieval operation performance, we have performed experiments with four image-scales φ , which correspond to 25%, 50%, 75% and 100% of the image size, i.e. $\varphi = \{0.25, 0.50, 0.75, 1\}$. For each of the views, our feature set is comprised by the feature vectors recovered by each of the three alternatives under study.

3.1 Columbia University COIL-20 Database

The COIL-20 database contains 72 views for 20 objects acquired by rotating the object under study about the vertical axis. The scaling of the views and this rotation account for the affine transformations mentioned earlier. In Figure 2, we show sample views for each of the objects in the database.

For our feature-based image retrieval experiments, we have removed 10 out of the 72 views for each object, i.e. every other seven views. These views are our query images. The views in the database constitute our data set, i.e.

20×62 views. For each of our query views, we retrieve the four images from the data set which amount to the highest values of the voting score β . Ideally, these scheme should select the four “data” views indexed immediately before and after the “query” view. In other words, the correct retrieval results for the “query” view indexed i are those views indexed $i - 2, i - 1, i + 1$ and $i + 2$. This scheme allows us to use the number of correctly recovered views as a measure of the accuracy of the matching algorithm and, hence, lends itself naturally to the performance assessment task in hand.

In Figure 3, we show, for each scale and image feature alternative, i.e. Harris corners, MSERs and SIFT descriptors, the mean retrieval rate as a function of object index. We also indicate, using error bars, the standard error for the ten query images per object. In the plots, we have used the indexing provided in Figure 2.

3.2 Urban-scene Database

Having presented evaluation results on the COIL-20 database, we now turn our attention to a more challenging setting. This is provided by a database of urban scenes. This database contains 110 views for 10 scenes acquired

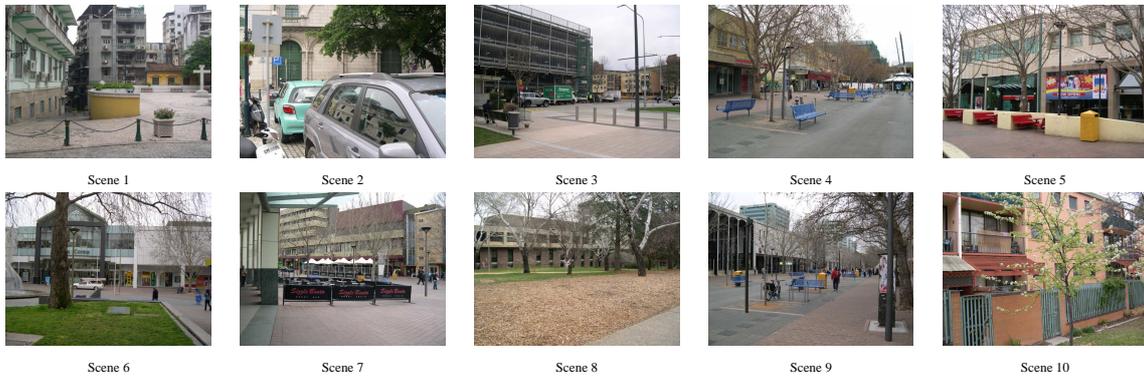


Figure 4: Sample views for the scenes in our database.

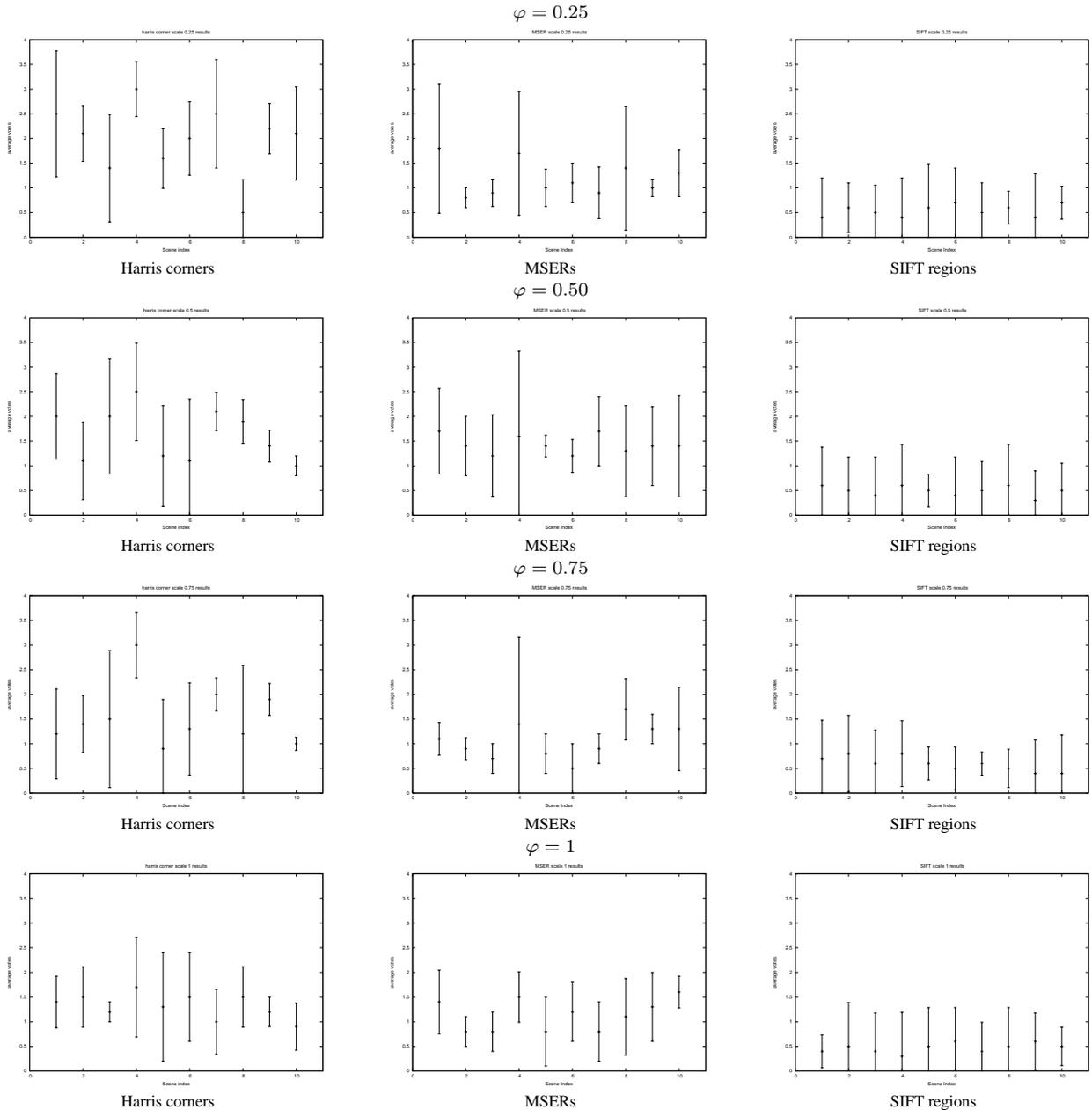


Figure 5: Retrieval results for the three alternatives of image feature vectors and four different values of the scaling factor φ .

by rotating the camera about its vertical axis from 0° to 66° degrees in steps of 6° , i.e. 11 views per scene. This viewpoint rotation, in conjunction with the scaling operations on the imagery, accounts for the affine transformations of the scene under study. In Figure 4, we show sample views for each of the scenes in the database.

For our feature-based image retrieval experiments, we

have followed an akin approach to that employed on the COIL-20 database. At this point, it is worth noting that, since our images are true-colour ones, we have converted them into gray-scale. After performing this conversion as a preprocessing step, we have removed 3 out of the 11 views for each object. This amounts to one every other three views. We use the excised views as query images,

whereas the remaining 80 images in the database constitutes our data set. As done previously, we retrieve the four images from the data set which amount to the highest values of the voting score β for each of the query views. Again, the correct retrieval results for the “query” images are those immediately before and after the view of reference. This scheme, being consistent with the one used to assess the performance of the image retrieval results on the COIL database, not only permits the direct association of the number of correctly recovered views to the accuracy measures computed from our experiments, but allows a direct comparison between the datasets used in both parts of our quantitative study.

In Figure 5, we repeat the sequence in Figure 3 for our urban-scene database. The plots show the mean and standard deviation for the retrieval rates as a function of object index and image-scale φ . In the plots, we have used the indexing provided in Figure 4.

3.3 Discussion

From the plots, we can conclude that the best performance, in terms of mean retrieval rate is given by the Harris corners. This is regardless of the scaling factor φ . Despite providing a margin of improvement in terms of performance with respect to the alternatives, the standard error for the Harris corners is the largest, as it is the variation in mean retrieval rate between scales. It is also worth noting that the SIFT is scale invariant, as claimed in (Lowe 2004). Thus, for the SIFT, the retrieval rate variation is small between scales, the retrieval rate itself is the lowest of the three alternatives. This may be due to the fact that the SIFT can be affected by viewpoint rotations. Finally, the MSERs show a sensitivity to scale and rotation which delivers a mean retrieval rate and standard error which are half-way between the Harris corners and the SIFT. This is in accordance with the stability assumption on the regions recovered by the algorithm in (Matas et al. 2002).

4 Conclusions

In this paper, we have presented an evaluation of three local image descriptors for purposes of contents-based image retrieval. In our experiments, we have accounted for the effects of rotation and scaling transformations on the retrieval rate and the standard error. Despite the evaluation presented here is not exhaustive, our experimental setting is quite general and can be easily extended to other descriptors elsewhere in the literature. Furthermore, the KD-Tree algorithm used here may be substituted with other relational matching algorithms for purposes of further comparison and evaluation.

References

Bentley, J. (1975), ‘Multidimensional binary search trees used for associative searching’, *Communications of the ACM* **8**(9).

Caneiro, G. & Jepson, A. (2002), Phase-based local features, in ‘European Conference on Computer Vision’, pp. I: 282–296.

Farzin, S. A. M. & Kittler, J. (1996), Robust and efficient shape indexing through curvature scale space, in ‘Proceedings of the 7th British Machine Vision Conference’, Vol. 1, pp. 53–62.

Fitzgibbon, A., Pilu, M. & Fisher, R. B. (1999), ‘Direct least square fitting of ellipses’, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **21**(5), 476–480.

Harris, C. J. & Stephens, M. (1988), A combined corner and edge detector, in ‘Proc. 4th Alvey Vision Conference’, pp. 147–151.

Lowe, D. (2004), ‘Distinctive image features from scale-invariant keypoints’, *International Journal of Computer Vision* **60**(2), 91–110.

Matas, J., Chum, O., Martin, U. & Pajdla, T. (2002), Robust wide baseline stereo from maximally stable extremal regions, in ‘Proceedings of the British Machine Vision Conference’, pp. 384–393.

Mikolajczyk, K. & Schmid, C. (2005), ‘A performance evaluation of local descriptors’, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **27**(10), 1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. & Gool, L. V. (n.d.), A comparison of affine region detectors. Submitted to the International Journal of Computer Vision, downloadable from “<http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05>”.

Niblack et al., W. (1993), The qbic project: Querying images by content using color, texture and shape, in ‘Proc. SPIE Conference on Storage and Retrieval of Image and Video Databases’, 1908, pp. 173–187.

Pentland, A. P., Picard, R. W. & Sclaroff, S. (1994), Photobook: tools for contents based manipulation of image databases, in ‘Storage and Retrieval for Image and Video Database’, Vol. II, pp. 34–47.

Picard, R. W. (1995), Light-years from lena: Video and image libraries ind the future, in ‘International Conference on Image Processing’, Vol. 1, pp. 310–313.

Randen, T. & Husoy, J. H. (1999), ‘Filtering for texture classification : A comparative study’, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **21**(4), 291–310.

Sclaroff, S. & Pentland, A. (1993), A model framework for correspondence and description, in ‘Proceedings of the International Conference on Computer Vision’, pp. 308–313.

Tamura, H. & Yokoya, N. (1984), ‘Image databe systems: a survey’, *Pattern Recognition* **17**(1), 29–49.

Varma, M. & Zisserman., A. (2003), Texture classification: Are filter banks necessary?, in ‘Conference on Computer Vision and Pattern Recognition’, pp. I:477–484.

Observer Annotation of Affective Display and Evaluation of Expressivity: Face vs. Face-and-Body

Hatice Gunes and Massimo Piccardi

Faculty of Information Technology, University of Technology, Sydney (UTS)
P.O. Box 123, Broadway 2007, NSW, Australia
{haticeg, massimo} @ it.uts.edu.au

Abstract

A first step in developing and testing a robust affective multimodal system is to obtain or access data representing human multimodal expressive behaviour. Collected affect data has to be further annotated in order to become usable for the automated systems. Most of the existing studies of emotion or affect annotation are monomodal. Instead, in this paper, we explore how independent human observers annotate affect display from monomodal face data compared to bimodal face-and-body data. To this aim we collected visual affect data by recording the face and face-and-body simultaneously. We then conducted a survey by asking human observers to view and label the face and face-and-body recordings separately. The results obtained show that in general, viewing face-and-body simultaneously helps with resolving the ambiguity in annotating emotional behaviours.

Keywords: Affective face-and-body display, bimodal affect annotation, expressivity evaluation.

1 Introduction

Affective computing aims to equip computing devices with the means to interpret and understand human emotions, moods, and possibly intentions without the user's conscious or intentional input of information—similar to the way that humans rely on their senses to assess each other's state of mind. Building systems that detect, understand, and respond to human emotions could make user experiences more efficient and amiable, customize experiences and optimize computer-learning applications.

Over the past 15 years, computer scientists have explored various methodologies to automate the process of emotion/affective state recognition. One major present limitation of affective computing is that most of the past research has focused on emotion recognition from one single sensorial source, or modality: the face (Pantic et

al., 2005). Relatively few works have focused on implementing emotion recognition systems using affective multimodal data (i.e. affective data from multiple channels/sensors/modalities). While it is true that the face is the main display of a human's affective state, other sources can improve the recognition accuracy. Emotion recognition via body movements and gestures has recently started attracting the attention of computer science and human-computer interaction (HCI) communities (Hudlicka, 2003). The interest is growing with works similar to these presented in (Balomenos et al., 2003), (Burgoon et al., 2005), (Gunes and Piccardi, 2005), (Kapoor and Picard, 2005) and (Martin et al., 2005).

A first step in developing and testing a robust affective multimodal system is to obtain or access data representing human multimodal expressive behaviour. The creation or collection of such data requires a major effort in the definition of representative behaviours, the choice of expressive modalities, and the labelling of large amount of data. At present publicly-available databases exist mainly for single expressive modalities such as facial expressions, static and dynamic hand postures, and dynamic hand gestures (Gunes and Piccardi, 2006b). Only recently, a first bimodal affect database consisting of expressive face and face-and-body display has been released (Gunes and Piccardi, 2006a).

Besides acquisition, another equally challenging procedure is their annotation. Multimodal data have to be annotated in order to become usable for the automated systems.

Most of the experimental research that studied emotional behaviours or affective data collection focused only on single modalities, either facial expression or body movement. In other words, the amount of information separate channels carry for recognition of emotions has been researched separately (explained in detail in Related Work section). There also exist several studies that involve multimodal annotation specific to emotions. However, none of the studies dealing with multimodal annotation specific to emotion compared how independent human observers' annotation is affected when they are exposed to a single modality versus multiple modalities occurring together. Therefore, in this paper, we conduct a study on whether seeing emotional displays from the face camera alone or from the face-and-body camera affects the independent observers' annotations of emotion. Our investigation

Copyright © 2006, Australian Computer Society, Inc. This paper appeared at the *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

focuses on evaluating monomodal versus bimodal posed affective data. Our aim is to use the annotations and results obtained from this study to train an automated system to support unassisted recognition of emotional states. However, creating, training and testing and affective multimodal system is not the focus of this paper.

2 Related Work

2.1 Emotion Research

In general, when annotating affect data two major studies from emotion research are used: Ekman's theory of emotion universality (Ekman, 2003) and Russell's theory of arousal and valence (Russell, 1980).

Ekman conducted various experiments on human judgement on still photographs of posed facial behaviour and concluded that seven basic emotions can be recognized universally, namely, neutral, happiness, sadness, surprise, fear, anger and disgust (Ekman, 2003). Several other emotions and many combinations of emotions have been studied but it remains unconfirmed whether they are universally distinguishable.

Other emotion researchers took the dimensional approach and viewed affective states not independent of one another; rather, related to one another in a systematic manner (Russell, 1980). Russell argued that emotion is best characterized in terms of a small number of latent dimensions, rather than in a small number of discrete emotion categories. Russell proposed that each of the basic emotions is a bipolar entity as part of the same emotional continuum. The proposed polarities are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). The model is illustrated in Figure 1.

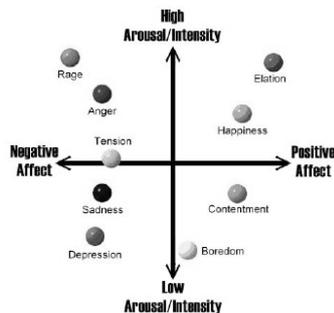


Figure 1. Illustration of Russell's circumplex model.

2.2 Affective multimodal data collection

All of the publicly available facial expression or body gesture databases collected data by instructing the subjects on how to perform the desired actions (please see (Gunes and Piccardi, 2006b) for an extensive review of publicly available visual affect databases).

2.3 Affective multimodal data annotation

Hereby, we review studies that deal with human annotation of non-verbal emotional behaviour. This review is intended to be illustrative rather than exhaustive. We do not review studies on human labelling and recognition of emotions from face expressions, as

they have been extensively reviewed by Ekman (Ekman, 1982; Ekman, 2003).

In (DeMeijer, 1991), the authors studied the attribution of aggression and grief to body movements. Three parameters in particular were investigated: sex of the mover, sex of the perceiver, and expressiveness of the movement. Videos of 96 different body movements from students of expressive dance were shown to 42 adults. The results showed that the observers used seven dimensions for describing movements: trunk movement (stretching, bowing), arm movement (opening, closing), vertical direction (upward, downward), sagittal direction (forward, backward), force (strong-light), velocity (fast-slow), directness (moving straight towards the end-position versus following a lingering, s-shaped pathway). The results of this study revealed that form and motion are relevant factors when decoding emotions from body movement.

In another study on bodily expression of emotion, Wallbott recorded acted body movements for basic emotions (Wallbott, 1998). Twelve drama students were then asked to code body movement and posture performed by actors. The results revealed that the following factors appeared to be significant in the coding procedure: position of face-and-body, position of shoulders, position of head, position of arms, position of hands, movement quality (movement activity, spatial expansion, movement dynamics, energy, and power); body movements (jerky and active), body posture.

In (Montepare et al., 1999), the authors conducted an experiment on the use of body movements and gestures as cues to emotions in younger and older adults. They first recorded actors doing various body movements. In order to draw the attention of the human observers to the expression of emotions via body cues, the authors electronically blurred the faces and did not record sound. In the first part of the experiments, the observers were asked to identify the emotions displayed by young adult actors. In the second part of the experiment, the observers were asked to rate the actors' displays using characteristics of movement quality (form, tempo, force, and direction) rated on a 7-point scale and verbal descriptors (smooth / jerky, stiff / loose, soft / hard, slow / fast, expanded / contracted, and no action / a lot of action). Overall, observers evaluated age, gender and race; hand position; gait; variations in movement form, tempo, and direction; and movement quality from actors' body movements. The ratings of both younger and older groups had high agreement when linking particular body cues to emotions.

Coulson presented experimental results on attribution of six emotions (anger, disgust, fear, happiness, sadness and surprise) to static body postures by using computer-generated figures (Coulson, 2004). He found out that in general, human recognition of emotion from posture is comparable to recognition from the voice, and some postures are recognized as well as facial expressions.

All of the aforementioned studies focused on individual modalities such as facial expression or body movement; therefore, in this paper we focus on bimodal data.

There exist several studies that involve multimodal annotation specific to emotions. The Belfast naturalistic

database contains emotional interviews annotated with continuous dimensions (Douglas-Cowie et al., 2003). In (Allwood et al., 2004) authors designed a coding scheme for the annotation of 3 videos of TV interviews. Facial displays, gestures, and speech were coded using the following parameters: form of the expression and of its semantic-pragmatic function (e.g. turn managing) and the relation between modalities: repetition, addition, substitution, contradiction. (Martin et al., 2005) also designed a coding scheme for annotating multimodal behaviours during real life mixed emotions. They first collected emotionally rich TV interviews. Then they focused on the annotation of emotion specific behaviours in speech, head and torso movements, facial expressions, gaze, and hand gestures. They grounded their coding scheme on the following parameters: the expressivity of movements, the number of annotations in each modality, their temporal features (duration, alternation, repetition, and structural descriptions of gestures), the directions of movements and the functional description of relevant gestures.

The materials collected by (Martin et al., 2005) are useful multimodal data for research in affective multimodal HCI. However, as annotation in itself is challenging and ambiguous, we believe that the annotation should be done more systematically than just one observer. Moreover, the annotation in (Martin et al., 2005) focused more on actions and expressions rather than emotions.

Therefore, in this paper, we explore whether seeing emotional displays from the face camera alone or from the face-and-body camera affects the independent observers' annotations of emotion.

3 Study

In this study we are seeking answers to the following research questions.

- How do humans perceive emotions from face modality alone compared to the combination of face-and-body modalities that occur simultaneously?
- Does being exposed to the expressions from one sensor (face camera only) or from multiple sensors simultaneously (viewing face-and-body combined) affect the observers' interpretations and therefore, labelling differ (monomodal vs. bimodal)?
- Does the use of multiple modalities help simplify the human affect recognition or on the contrary makes it more complicated? Does it help with resolving ambiguity or the addition of another modality increases ambiguity?

3.1 The data set

The data set we used for this study consists of recordings of combined face and body expressions. According to five factors that were defined by Picard in (Picard et al., 2001) as influencing the affective data collection, the data we collected are: posed, obtained in laboratory settings, with an emphasis on expression rather than feelings, openly recorded and obtained with an emotion purpose. This is consistent with the characteristics of most of the available face and body gesture databases (Gunes and Piccardi, 2006b).

We recorded the video sequences simultaneously using two fixed cameras with a simple setup and uniform background. One camera was placed to specifically capture the face only and the second camera was placed in order to capture face-and-body movement from the waist above. Prior to recordings subjects were instructed to take a neutral position, facing the camera and looking straight to it with hands visible and placed on the table. The subjects were asked to perform face and body gestures simultaneously by looking at the facial camera constantly. The recordings were obtained by using a scenario approach that was also used in previous emotion research (e.g. Wallbott and Scherer, 1986). In this approach, subjects are provided with situation vignettes or short scenarios describing an emotion eliciting situation. They are instructed to imagine these situations and act out as if they were in such a situation. In our case the subjects were asked what they would do when "it was just announced that they won the biggest prize in lottery" or "the lecture is the most boring one and they can't listen to it anymore" etc. In some cases the subjects came up with a variety of combinations of face and body gestures. As a result of the feedback and suggestions obtained from the subjects, the number and combination of face and body gestures performed by each subject varies slightly (see (Gunes and Piccardi, 2006a) for details). Fig. 2 shows representative images obtained simultaneously by the body and face cameras.

3.2 The annotation method

Once the multimodal data are acquired, they need to be annotated and analysed to form the ground truth for machine understanding of the human affective multimodal behaviour. Annotation of the data in a bimodal/multi modal database is a very tiresome procedure overall as it requires extra effort and time to view and label the sequences with a consistent level of alertness and interest. Hence, obtaining the emotion- and quality-coding for all the visual data contained in bimodal databases is extremely tedious and very difficult to achieve.

We obtained the annotation of our visual multimodal data (each face and body video separately) by asking human observers to view and label the videos. The purpose of this annotation was to obtain independent interpretations of the displayed face and body expressions; evaluate the performance (i.e. how well the subjects were displaying the affect they intended to communicate using their face and bodily gesture) by few human observers from different ethnic and/or cultural background.

To this aim, we developed a survey for face and body videos separately, using the labelling schemes for emotion content and signs. We used two main labelling schemes in line with the psychological literature on descriptors of emotion: (a) verbal categorical labelling (perceptually determined, i.e. happiness) in accordance with Ekman's theory of emotion universality (Ekman, 2003) and (b) broad dimensional labelling: arousal/activation (arousal-sleep/ activated -deactivated) in accordance with Russell's theory of arousal and valence (Russell, 1980). The participants were first shown the whole set of facial videos and only after

finishing with the face they were shown the corresponding body videos. For each video they were asked to choose one label only, from the list provided: sadness, puzzlement/thinking, uncertainty/"I don't know", boredom, neutral surprise, positive surprise, negative surprise, anxiety, anger, disgust, fear, happiness.

For the face videos the procedure was as follows. We asked each participant to select labels for the numbered videos they were shown. When they had difficulty choosing a label they were encouraged to guess. Secondly, we asked each participant to choose a number between 1 and 10 as to how well the emotion is displayed (1 indicating "low" and 10 indicating "high" quality in the expressiveness).

For the body videos the procedure was as follows. We asked each participant to select labels for the numbered videos they were shown. When they had difficulty choosing a label again they were encouraged to guess. Secondly, we asked each participant to choose a number between 1 and 10 as to (a) how fast or slow the motion occurs in the display (i.e. movement speed): 1 indicating "very slow" and 10 indicating "very fast"; (b) how the movement causes the body's occupation of space in the display (i.e. movement in space): 1 indicating "very contracted/very less space coverage" and 10 indicating "very expanded/a lot of space coverage" during the movement; and (c) how powerful/energetic the movement displayed is (i.e. movement dynamics): 1 indicating "almost no action" and 10 indicating "a lot of action" in the movement.

360 face and 360 face-and-body videos were annotated in this way and results analysed.

3.3 Participants

We chose videos from 15 subjects and divided them based on the subjects into three sub-sets to make the annotation procedure easier. Eventually, the first sub-set contained 124 face and 124 body videos from five subjects and was viewed and annotated by six observers: Bulgaria (1), Turkey (1), Mexico (1), Pakistan (1), Czech Republic (1), and Australia (1). The second sub-set contained 120 face and 120 body videos from other five subjects and was viewed and annotated by six observers. Observers were from the following countries: Bulgaria (1), Turkey (1), Czech Republic (2), Slovakia (1), and China (1). The third sub-set contained 116 face and 116 body videos from other five subjects and was viewed and annotated by six observers: Bulgaria (1), Turkey (1), Czech Republic (2), Brazil (1), and China (1).

3.4 Results

For each video, all labelling provided by the six observers was analysed and the emotion category that received the highest vote as unique was used to finalize the true label of that particular video, thus, the ground truth. The display from certain subjects can be classified to a particular emotion category almost unambiguously (i.e. all six observers agree that the display is of one particular emotion category), which implies that these actors produced rather stereotyped movements

irrespective of the emotion to be encoded. The classification results for other actors are observed to be more ambiguous (i.e. not all six observers agree that the display is of one particular emotion category). For face videos, "quality of expressiveness" was obtained by averaging the six quality votes provided by the observers. For body videos, results for "movement speed", "movement in space", and "movement dynamics", were similarly obtained by averaging the six votes provided by the observers.

According to the results obtained from both face and face-and-body video annotation:

- 295 out of 360 videos were labelled using the same emotion label both for the face videos and for the face-and-body videos. 65 videos were labelled differently.
- 140 out of 360 videos have more agreement for the face-and-body video than the face video alone.
- 125 out of 360 videos have same level of agreement for the face-and-body video and the face video alone.
- 95 out of 360 videos have more agreement for the face video only than the face-and-body video.

3.4.1 Results for Face Videos

The details of the independent observer agreement for face videos are presented in Table 1.

critierion	# of videos for face	# of videos for face-and-body
Higher than 3 votes	292	311
Higher than 4 votes	200	234
Higher than 5 votes	114	139
Equal to 6 votes	84	118

Table 1. The details of the independent observer agreement for face and face-and-body videos: number of videos complying with the criterion.

The emotion categories that caused more cross-confusion when labelling the face data are *puzzlement* and *anxiety*. This can be due to the fact that both emotions were expressed with similar facial displays (e.g. lip bite). Viewing face-and-body display together almost immediately helped the observers resolve their ambiguity. This in turn suggests that if physical displays for certain emotions are similar, and no specific, discriminative movement indicators exist, in independent observer labelling, these emotion displays are commonly found to be confused with one another.

When expressivity of the face videos was analysed it was found that the videos that did not have high agreement in terms of emotion labelling not necessarily were rated low in terms of expressivity. In other words, an observer rated the expressivity of the face display assuming that the person was expressing the emotion s(he) thought was the true emotion displayed.

3.4.2 Results for the Combined Face-and-Body Videos

The details of the independent observer agreement for face-and-body videos are presented in Table 1. Further results from the face-and-body video annotation are presented in Tables 2-4.

According to the results presented in Table 1 we conclude that full agreement is achieved more frequently when face-and-body are viewed together (118 compared

to 84). The results provided in Table 2 and Table 3 suggest that the emotion with lowest movement speed, least movement in space and least movement dynamics in space are *sadness*, followed by *puzzlement*, *anxiety*, and *uncertainty*.

Emotion	Average movement speed	Average movement in space	Average movement dynamics
sadness	3.89	4.63	5.11
puzzlement	4.23	4.55	4.96
uncertainty	4.55	4.53	4.72
boredom	4.56	4.80	5.25
surprise	4.83	5.01	5.29
anxiety	4.98	3.73	4.83
anger	5.32	4.76	5.46
disgust	5.41	4.62	5.61
fear	6.05	4.85	6.21
happiness	6.13	5.54	6.32

Table 2. The details of the face-and-body survey: the average movement speed, average movement in space and average movement dynamics criteria for each emotion category.

Order	average movement speed	average movement in space	in	average movement dynamics
1	happiness	happiness		happiness
2	fear	surprise		fear
3	disgust	fear		disgust
4	anger	boredom		anger
5	anxiety	anger		surprise
6	surprise	sadness		boredom
7	boredom	disgust		sadness
8	uncertainty	puzzlement		puzzlement
9	puzzlement	uncertainty		anxiety
10	sadness	anxiety		uncertainty

Table 3. The details of the face-and-body survey: Ranking of the emotion categories (in descending order) based on the average movement speed, average movement in space and average movement dynamics criteria.

These emotion categories fall in the “low intensity/arousal” category in Russell’s circumflex model. The emotion with highest movement speed, largest movement in space and highest movement dynamics in space are *happiness*, followed by *fear*, *surprise* and *disgust*. These emotion categories fall in the “high intensity/arousal” category in Russell’s circumflex model (see Fig. 1) (Russell, 1980).

According to the results compiled in Table 4, we can state that bimodal data helps with resolving ambiguity in most of the cases. The usefulness of the bimodal data for observer annotation is two-fold: (a) resolving ambiguity and (b) re-labelling of the videos.

(a) *Resolving ambiguity that is present in affect annotation of the face data*

Of the 65 videos that were labelled differently, ambiguity was resolved for 27 videos using the face-and-body data. This fact can be illustrated with the following examples:

- A face video that obtained divided votes by the observers (3 boredom and 3 puzzlement) was later labelled as *boredom* with much more certainty (5 votes) (video # S001-012).

- A face video that obtained divided votes by the observers (3 puzzlement and 3 anxiety) was later labelled as *anxiety* by all 6 observers (video # S001-040, see Fig. 2, left hand side).
- A face video that obtained divided votes by the observers (1 anger, 1 puzzlement 2 sadness, 1 ambiguity, 1 boredom), when viewed with face and body together, was labelled as *anxiety* by 5 observers (video # S002-010).
- A face video that obtained divided votes by the observers (3 boredom and 3 sadness), when viewed with face and body together, was labelled as *boredom* by 4 observers (video # S002-011).
- A face video that obtained divided votes by the observers (3 boredom and 3 sadness), when viewed with face and body together, was labelled as *anxiety* by all 6 observers (video # S010-039, see Fig. 2, right hand side).

(b) *Changing the label of the displayed emotion obtained from face data to another label*

Of the 65 videos that were labelled differently, 19 videos were re-labelled with *almost always higher agreement* when face-and-body data was viewed. This fact can be illustrated with the following examples:

- A face video that was labelled as *negative surprise*, when viewed as face and body together, was labelled as *positive surprise* (video # S001-007).
- A face video that was labelled as *puzzlement* by the observers (4 votes out of 6), when viewed as face and body together, was labelled as *anxiety* by the all 6 observers (video # S001-043).

In one case (video # S013-018), the face-and-body data helps with decreasing the level of ambiguity, but is not sufficient to resolve it. However, in 18 cases (see Table 3) the body adds ambiguity to the annotation. According to the results presented in Table 4, the emotion categories that caused more confusion in the bimodal data are *happiness* and *positive surprise* (7 out of 18 cases). Happiness was expressed as *extreme joy* and some observers labelled this display as *positive surprise*, which in fact is not wrong.

4 Discussion and Conclusions

Our investigation focused on evaluating monomodal and bimodal posed affective data for the purpose of aiding multimodal affect recognition systems that are dependent on human affective state as their input for interaction. According to the results obtained we conclude that in general, bimodal face-and-body data helps with resolving ambiguity carried by the face data alone. This in turn suggests that an automatic multimodal affect recognizer should attempt to combine facial expression and body gestures for improved recognition results.

video #	label for face video	votes	label for the combined face-and-body video	votes	changes & interpretation
s001-07	negative surprise	4	positive surprise	5	label changed and higher level of agreement between observers
s001-12	boredom- puzzlement	3—3	boredom	5	ambiguity resolved, label changed and higher level of agreement between observers
s001-23	fear- negative surprise	3—3	fear	5	ambiguity resolved, label changed and higher level of agreement between observers
s001-40	puzzlement- anxiety	3—3	anxiety	6	ambiguity resolved, label changed and full agreement between observers
s001-42	puzzlement- anxiety	3—3	anxiety	6	ambiguity resolved, label changed and full agreement between observers
s001-43	puzzlement	4	anxiety	6	label changed and full agreement between observers
s001-44	puzzlement	3	anxiety	6	label changed and full agreement between observers
s002-01	happiness- positive surprise	3—3	happiness	4	ambiguity resolved, label changed and higher level of agreement between observers
s002-10	sadness	2	anxiety	5	label changed and higher level of agreement between observers
s002-11	boredom-sadness	3—3	boredom	4	ambiguity resolved, label changed and higher level of agreement between observers
s003-01	negative surprise	2	happiness	3	label changed and higher level of agreement between observers
s003-05	puzzlement	4	uncertainty-puzzlement	2—2	bimodal data causes ambiguity
s003-08	boredom-puzzlement	2—2	boredom	4	ambiguity resolved, label changed and higher level of agreement between observers
s003-11	boredom	3	boredom-anxiety	3—3	bimodal data causes ambiguity
s004-19	puzzlement	3	anxiety	3	label changed
s005-07	uncertainty	3	anger	3	label changed
s005-14	puzzlement- anxiety- uncertainty	2—2—2	puzzlement	4	ambiguity resolved, label changed and higher level of agreement between observers
s005-22	boredom-puzzlement	3—3	boredom	5	ambiguity resolved, label changed and higher level of agreement between observers
s005-24	disgust	4	disgust-fear	3—3	bimodal data causes ambiguity
s005-32	negative surprise	5	negative surprise- neutral surprise	3—3	bimodal data causes ambiguity
s006-04	negative surprise-fear	3—3	negative surprise	3	label changed
s006-27	uncertainty- puzzlement	3—3	puzzlement	4	ambiguity resolved, label changed and higher level of agreement between observers
s006-29	sadness-boredom	2—2	sadness	3	ambiguity resolved, label changed and higher level of agreement between observers
s006-32	puzzlement	3	uncertainty	3	label changed
s008-02	neutral surprise	3	negative surprise- neutral surprise	3—3	bimodal data causes ambiguity
s008-05	happiness	4	happiness-positive surprise	3—3	bimodal data causes ambiguity
s008-07	puzzlement	3	boredom	6	label changed and full agreement between observers
s009-03	puzzlement	5	uncertainty	3	label changed
s009-12	puzzlement-sadness	3—3	puzzlement	4	ambiguity resolved, label changed and higher level of agreement between observers
s009-14	puzzled-anxiety	3—3	anxiety	4	ambiguity resolved, label changed and higher level of agreement between observers
s010-02	happiness	5	happiness-positive surprise	3—3	bimodal data causes ambiguity
s010-21	puzzlement-anxiety	3—3	puzzlement	5	ambiguity resolved, label changed and higher level of agreement between observers
s010-39	sadness-boredom	3—3	anxiety	6	ambiguity resolved, label changed and full agreement between observers
s010-42	negative surprise	5	negative surprise-fear	2—2	bimodal data causes ambiguity
s011-01	happiness	6	happiness-positive surprise	3—3	bimodal data causes ambiguity
s011-02	happiness	5	happiness-positive surprise	3—3	bimodal data causes ambiguity
s011-03	anger	3	negative surprise-anger	3—3	bimodal data causes ambiguity
s011-13	puzzlement	4	uncertainty-puzzlement	2—2	bimodal data causes ambiguity
s011-15	boredom	4	puzzlement	4	label changed
s011-24	puzzlement	2	anxiety-boredom- puzzlement	2—2—2	bimodal data causes ambiguity
s012-01	happiness	6	positive surprise- happiness	3—3	bimodal data causes ambiguity
s012-05	neutral surprise	3	anger	4	label changed and higher agreement between observers
s012-14	fear-negative surprise	3—3	fear	4	ambiguity resolved, label changed and higher level of agreement between observers
s012-20	anxiety-puzzlement	2—2	puzzlement	3	ambiguity resolved, label changed and higher level of agreement between observers
s013-01	happiness	5	positive surprise- happiness	3—3	bimodal data causes ambiguity
s013-03	sadness	4	anger	5	label changed and higher agreement between observers
s013-12	happiness-positive surprise	3—3	positive surprise	4	ambiguity resolved, label changed and higher level of agreement between observers
s013-15	fear-disgust	2—2	fear	5	ambiguity resolved, label changed and higher level of agreement between observers
s013-18	anxiety-boredom- uncertainty	2—2—2	anxiety-boredom	3—3	higher level of agreement, however ambiguity between two labels still exists
s014-01	happiness	5	happiness-positive surprise	3—3	bimodal data causes ambiguity
s014-02	puzzlement- uncertainty	3—3	uncertainty	4	ambiguity resolved, label changed and higher level of agreement between observers
s014-03	anger-negative surprise	2—2	anger	4	ambiguity resolved, label changed and higher level of agreement between observers
s014-06	sadness-negative surprise	3—3	negative surprise	3	ambiguity resolved and label changed
s014-07	sadness-anxiety	2—2	anxiety	3	ambiguity resolved, label changed and higher level of agreement between observers
s015-09	anger-disgust	2—2	anger	3	ambiguity resolved, label changed and higher level of agreement between observers
s015-12	uncertainty	3	puzzlement	3	label changed
s015-14	puzzlement	3	boredom-uncertainty- puzzlement	2—2—2	bimodal data causes ambiguity
s015-17	sadness-boredom	2—2	puzzlement	3	label changed and higher agreement between observers
s015-19	puzzlement	3	boredom	3	label changed
s015-22	happiness	4	positive surprise	5	label changed and higher agreement between observers

s015-28	sadness-boredom	3—3	boredom	5	ambiguity resolved, label changed and higher level of agreement between observers
s016-03	neutral surprise	3	positive surprise	4	label changed and higher agreement between observers
s016-04	neutral surprise	2	positive surprise	3	label changed and higher agreement between observers
s016-05	anger	3	anger-positive surprise	2—2	bimodal data causes ambiguity
s016-11	puzzlement	3	boredom	3	label changed

Table 4. List of the videos that were labelled differently for face and face-and-body modalities and the details of the labelling results.

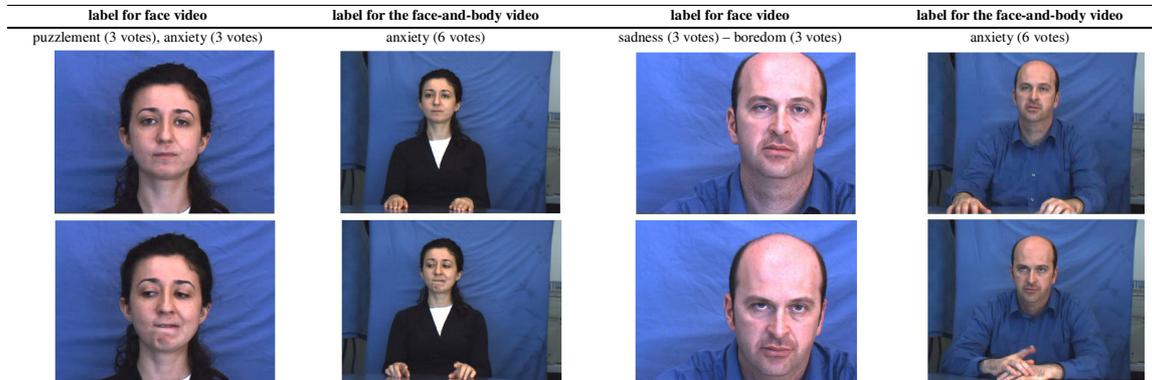


Figure 2. Example videos that were annotated differently for face and face-and-body and labels obtained from the survey: video # s001-40 (left hand side) and video # s010-039 (right hand side); neutral frames (first rows), expressive frames (second rows).

Overall, from the results obtained we can state that during annotation only in seldom cases do six observers fully agree on the emotion labelling. However, in general there is substantial agreement between the observers.

Affective state annotation in itself faces three main challenges (a) the type of emotion encoded, (b) the specific ability of the encoder, and (c) specific, discriminative movement indicators for certain emotions versus indicators of the general intensity of the emotional experience (Wallbott, 1998). Moreover, for the annotation purposes it is almost impossible to use emotion words that are agreed upon by everybody. The problem of what different emotion words are used to refer to the same emotion display is not, of course, a problem that is unique to this; it is by itself a topic of research for emotion theorists and psychologists. It is a problem deriving from the vagueness of language, especially with respect to terms that refer to psychological states (Ortony and Turner, 1990).

Furthermore, it is arguable that there may be differences in interpretation of the annotation scheme used to scale the expressivity of face and body. According to the results obtained we conclude that in general independent human observers tend to give average marks (i.e. 4 – 6 over a scale of 10) when rating speed, space usage and movement dynamics of the affective body movement. These results might be explained by the fact that there are some inherent difficulties in marking schemes in general (Blumhof and Stallibrass, 1994). These difficulties include:

- tendency to mark the more immediate concepts;
- tendency to mark towards the middle;
- exposing the subjectivity of marking schemes by trying to decide on, and weight, criteria. For instance, a mark of seven might represent a high mark for one observer,

whereas the same mark for another observer might represent a concept of just above average.

One major finding of this study is the fact that bimodal data helps with resolving ambiguity in most of the cases (46 out of 65). However, in 18 cases (see Table 4) the body adds ambiguity to the recognition. The strategy to follow in such cases could be to ask an additional group of observers to view and label the data.

Our analysis suggests that affective information carried by the bimodal data is valuable and will aid an automatic multimodal affect recognizer achieve improved recognition results.

The relative weight given to facial expression, speech, and body cues depend both on the judgment task (i.e. what is rated and labelled) and the conditions in which the behaviour occurred (i.e. how the subjects were simulated to produce the expression) (Ekman, 1982). Despite many findings in emotion behaviour research, there is no evidence in the actual human-to-human interaction on how people attend to the various communicative channels (speech, face, body etc.). Assuming that people judge these channels separately or the information conveyed by these channels is simply additive, is misleading (Sebe et al., 2005). As future work, a study exploring these factors can be conducted.

In this study we recorded face and upper body using separate cameras, obtaining higher resolution for the face images and lower resolution for the upper body images. We did not analyse whether or not resolution poses a challenge for visual affect data interpretation and annotation. It is possible to further compare whether being exposed to face display with low resolution, face display with high resolution, and finally combined face-and-body display affects the human attention and perception of affective video data.

The experiment presented in this paper can further be extended by data obtained in natural and realistic settings.

As confirmed by many researchers in the field, directed affective face and body action tasks differ in appearance and timing from spontaneously occurring behaviour (Cohn et al., 2004). Deliberate face or body behaviour is mediated by separate motor pathways and differences between spontaneous and deliberate actions may be significant. However, collecting spontaneous multimodal affect data is a very challenging task involving ethical and privacy concerns together with technical difficulties (high resolution, illumination, multiple sensors, consistency, repeatability etc.). The research field of multimodal affective HCI is relatively new and future efforts have to follow (Pantic et al., 2005).

5 Acknowledgments

We would like to thank Aysel Gunes (Sydney Central College) for her help with the FABO recordings and the annotation procedure. We would also like to thank Michelle Murch (Production Support Coordinator, Film and Video Media Centre, Faculty of Humanities and Social Sciences, UTS) for her support regarding the technical issues for the video recordings, the anonymous participants for taking part in the recordings, and the anonymous observers for viewing and labelling the videos.

6 References

- Allwood, J. et al. (2004), 'The MUMIN multimodal coding scheme'. in *Proc. Workshop on Multimodal Corpora and Annotation*, Stockholm.
- Balomenos, T., et al. (2004), 'Emotion analysis in man-machine interaction systems', in *Proc. MLMI*, LNCS 3361, pp. 318–328.
- Blumhof, J., Stallibrass, C. (1994), 'Peer Assessment', Hatfield: University of Herefordshire.
- Burgoon, J. K., et al. (2005), 'Augmenting human identification of emotional states in video', in *Proc. Int. Conf. on Intelligent Data Analysis*.
- Cohn, J.F. et al. (2004), 'Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles', in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 129–135.
- Coulson, M. (2004), 'Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence', *J. of Nonverbal Behaviour*, Vol. 28, pp. 117–139.
- DeMeijer, M. (1991), 'The attribution of aggression and grief to body movements: the effect of sex-stereotypes'. *European Journal of Social Psychology*, Vol. 21.
- Douglas-Cowie, E. et al. (2003), 'Emotional speech: Towards a new generation of databases'. *Speech Communication*, Vol. 40.
- Ekman, P. (1982): *Emotions in the human faces*, 2 ed., Studies in Emotion and Social Interaction, Cambridge University Press.
- Ekman, P. (2003): *Emotions revealed*. Weidenfeld & Nicolson.
- Gunes, H. and Piccardi, M. (2005), 'Fusing Face and Body Display for Bi-Modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration', in *Proc. ACII*, LNCS 3784, pp. 102–111.
- Gunes, H. and Piccardi, M. (2006a), 'A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behaviour', in *Proc. ICPR*, Vol. 1, pp. 1148–1153.
- Gunes, H. and Piccardi, M. (2006b), 'Creating and Annotating Affect Databases from Face and Body Display: A Contemporary Survey', in *Proc. IEEE SMC* (in press).
- Hudlicka, E. (2003), 'To feel or not to feel: The role of affect in human computer interaction', *Int. J. Hum.-Comput. Stud.*, Vol. 59, No. (1–2), pp. 1–32.
- Kapoor, A. and Picard, R. W. (2005), 'Multimodal affect recognition in learning environments', in *Proc. ACM Multimedia*, pp. 677–682.
- Martin, J.C., Abrilian, S. and Devillers, L. (2005), 'Annotating Multimodal Behaviours Occurring During Non Basic Emotions', in *Proc. ACII*, LNCS 3784, pp. 550–557.
- Montepare, J. et al. (1999), 'The use of body movements and gestures as cues to emotions in younger and older adults', *Journal of Nonverbal Behaviour*, Vol. 23, No. 2.
- Ortony, A. and Turner, T. J. (1990), 'What's basic about basic emotions?', *Psychological Review*, Vol. 97, pp. 315–331.
- Pantic, M. et al. (2005), 'Affective multimodal human-computer interaction', in *Proc. ACM Multimedia*, pp. 669–676.
- Picard, R. W., Vyzas, E. and Healey, J. (2001), 'Toward Machine Emotional Intelligence: Analysis of Affective Physiological State', *IEEE Tran. PAMI*, Vol. 23, No. 10, pp. 1175–1191.
- Russell, J. A. (1980), 'A circumflex model of affect', *Journal of Personality and Social Psychology*, Vol. 39, pp. 1161–1178.
- Sebe, N., Cohen, I. and Huang, T.S. (2005), 'Multimodal emotion recognition', *Handbook of Pattern Recognition and Computer Vision*, World Scientific.
- Wallbott, H. G. and Scherer, K. R. (1986), 'Cues and channels in emotion recognition', *Journal of Personality and Social Psychology*, Vol. 51, pp. 690–699.
- Wallbott, H. G. (1998), 'Bodily expression of emotion', *European Journal of Social Psychology*, Vol. 2.

Face Refinement through a Gradient Descent Alignment Approach

Simon Lucey, Iain Matthews

Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: slucey@ieee.org, iainm@cs.cmu.edu

Abstract

The accurate alignment of faces is essential to almost all automatic tasks involving face analysis. A common paradigm employed for this task is to exhaustively evaluate a face template/classifier across a discrete set of alignments (typically translation and scale). This strategy, provided the template/classifier has been trained appropriately, can give one a reliable but “rough” estimate of where the face is actually located. However, this estimate is often too poor to be of use in most face analysis applications (e.g. face recognition, audio-visual speech recognition, expression recognition, etc.). In this paper we present an approach that is able to *refine* this initial rough alignment using a gradient descent approach, so as to gain adequate alignment. Specifically, we propose an efficient algorithm which we refer to as the *sequential algorithm*, which is able to obtain a good balance between alignment accuracy and computational efficiency. Experiments are conducted on frontal and non-frontal faces.

Keywords: Face Alignment, Gradient Descent Object Alignment, Inverse Compositional Algorithm.

1 Introduction

Discriminative classifiers have been used with great success in the area of object detection. Most of these approaches, however, have concentrated on simply training a classifier with positive (i.e., aligned) and negative (i.e., not aligned) example images of the object. This classifier is then used to detect an object in a given image by exhaustively searching through all possible translations and scales. The now popular work (Viola & Jones 2001) of Viola and Jones is a prime example of this type of approach to object alignment. Such approaches are useful for obtaining a rough estimate of where the object is in an image, but struggle when one requires an alignment with more degrees of freedom than just translation and scale; such as an affine warp.

Another option after applying an exhaustive face detector is to exhaustively search for descriptors within the object that are largely invariant to affine variations. Typically in frontal face detection the eye region has been used in this capacity to great effect. Notable examples of these type of approaches have been (Moghaddam & Pentland 1997, Everingham & Zisserman 2006, Rurainsky & Eisert 2004, Wang &

Ji 2005, Lowe 1999). A criticism of this approach, however, is that the availability of these invariant descriptors is not always assured and is very specific to the object being aligned. For example, faces undergoing view-point change often change appearance dramatically requiring the selection of different descriptors. The selection of these affine invariant descriptors is often based on heuristics, and it is still largely an open question how these descriptors change across view-point.

Gradient descent methods for object alignment, such as the Lucas-Kanade (LK) (Lucas & Kanade 1981) and the Inverse-Compositional (IC) (Baker & Matthews 2001) algorithms, provide a natural solution to these dilemmas for two reasons. First, they attempt to find a gradient descent solution to the optimal object alignment without having to resort to an impractical exhaustive search. Second, they also provide the desirable property of treating all objects in a unified way, thus not requiring any heuristically chosen affine invariant descriptors to be selected (e.g., eye detectors). A problem, however, is that gradient descent approaches have poor generalization properties when they have to deal with previously unseen intra-class object variation (Gross, Baker & Matthews 2005). A prime example of this problem is when one is trying to align a previously unseen face given one has a rough idea of where the face is located. A graphical depiction of this task can be seen in Figure 1. Unfortunately, discriminative learning methods cannot be as freely applied with gradient descent approaches as with exhaustive methods. Since gradient descent methods are inherently iterative one cannot treat misaligned images as just negative examples. Such images may be part of the solution trajectory. Inhibiting these images may actually stop the algorithm progressing towards the correct alignment.

The problem of dealing with appearance variation in gradient descent object alignment is not new. Most notably, Black and Jepson addressed the problem for the case of general appearance variation (Black & Jepson 1998). Similarly, Hager and Belhumeur conducted work for the more specific case of illumination variation (Hager & Belhumeur 1998). In recent work (Baker, Gross & Matthews 2003), Baker et al. presented a unifying framework in which much of this previous work could be subsumed. Additionally, Baker et al. proposed two broad strategies for dealing with appearance variation when performing gradient descent object alignment, specifically the *simultaneous* and *project-out* algorithms. The simultaneous algorithm is able to give good alignment accuracy, but is computationally slow due to the large matrix inversions that must be performed at each iteration. Conversely, the project-out algorithm is computationally fast, due to simplifying assumptions, but suffers from poorer alignment¹ performance.

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

¹We would like to note that the project-out algorithm has been

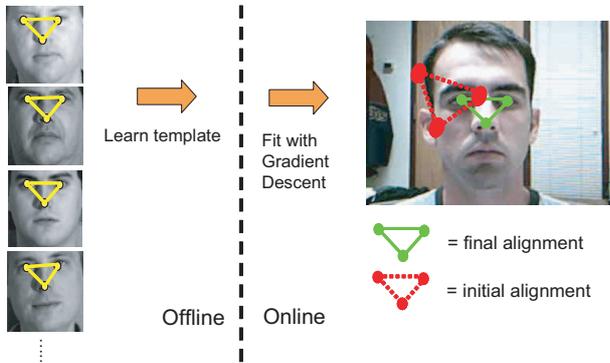


Figure 1: This figure depicts the task we want to undertake in this paper, where we have a rough approximation to where an object is (for our work in this paper the object will be a face), and we want to improve this alignment through a gradient descent fit. Faces naturally contain appearance variation, so we need to generalize from an offline ensemble of aligned face images so as to align to previously unseen subjects.

In this paper we propose a new algorithm that is able to give good alignment accuracy with reasonable computational efficiency. We refer to this approach as the *sequential algorithm*. The task we use throughout this paper to evaluate these approaches is face alignment, where we want to be able to accurately align, using an affine warp, a subject independent template to all faces; even if that face has not been previously seen offline. Experiments are conducted on frontal and non-frontal faces, demonstrating large improvements in alignment over canonical approaches.

2 The Inverse Compositional Algorithm

The Lucas-Kanade (LK) algorithm (Lucas & Kanade 1981) has become a common tool in computer vision for the task of image alignment. The inverse compositional (IC) image alignment algorithm (Baker & Matthews 2001), developed by Baker and Matthews, is a more efficient formulation of the LK algorithm. In the IC formulation many computationally costly components of the algorithm can be pre-computed from the template, unlike the LK algorithm. The IC algorithm is essentially the minimization of the following with respect to $\Delta \mathbf{p}$,

$$\|\mathbf{y}^{(\mathbf{p})} - \mathbf{t}^{(0)} - \mathbf{J}(\mathbf{t}^{(0)})\Delta \mathbf{p}\|^2 \quad (1)$$

where $\mathbf{y}^{(\mathbf{p})}$ is the vectorized form of the image $Y(\mathcal{W}(\mathbf{x}, \mathbf{p}))$, and $\mathbf{t}^{(0)}$ is the vectorized image of $T(\mathcal{W}(\mathbf{x}, \mathbf{0}))$ which is an approximation to the aligned image $Y(\mathcal{W}(\mathbf{x}, \mathbf{p}^*))$. An alignment function $\mathcal{W}(\mathbf{x}, \mathbf{p})$ is employed to map an image position \mathbf{x} to a new position \mathbf{x}' based on the warp parameters \mathbf{p} , where \mathbf{p}^* is the correct alignment we are attempting to estimate. Since the warp $\mathcal{W}(\mathbf{x}, \mathbf{p})$ is non-linear we must approximate it using the linear matrix,

shown (Baker et al. 2003) to perform very well in situations where the appearance variation has been previously seen offline, and that the rank of this variation is small. In this paper we are investigating the more general case where the appearance variation has not been seen previously, and the rank of this variation is quite large.

$$\mathbf{J}(\mathbf{t}) = [\nabla T(\mathcal{W}([0, 0]^T, \mathbf{0})) \frac{\partial \mathcal{W}}{\partial \mathbf{p}}, \dots, \nabla T(\mathcal{W}([N-1, M-1]^T, \mathbf{0})) \frac{\partial \mathcal{W}}{\partial \mathbf{p}}]^T \quad (2)$$

where the image T is an $N \times M$ image. For ease of notation \mathbf{t} was used in Equation 2, rather than $\mathbf{t}^{(0)}$ because the $(\mathbf{0})$ represents the identity warp $\mathcal{W}(\mathbf{x}, \mathbf{0})$; this convention will be used throughout the rest of this paper. One can see the approximation being used in Equation 2 is a first order Taylor's series approximation to the warp.

It is easy to show that the solution to Equation 1 is,

$$\Delta \mathbf{p} = (\mathbf{J}(\mathbf{t})^T \mathbf{J}(\mathbf{t}))^{-1} \mathbf{J}(\mathbf{t})^T (\mathbf{y}^{(\mathbf{p})} - \mathbf{t}) \quad (3)$$

Due to the linear approximation in Equation 2 this solution is not explicit so we must iterate until we get convergence. This form of optimization is commonly referred to as Gauss-Newton optimization. The warp parameter \mathbf{p} corresponds to the current estimate of the set of warp parameters needed to bring the two images into alignment, and $\Delta \mathbf{p}$ is the warp update that will improve the alignment. One can then update the warp estimate as follows:

$$\mathcal{W}(\mathbf{x}; \mathbf{p}) \leftarrow \mathcal{W}(\mathbf{x}; \mathbf{p}) \circ \mathcal{W}(\mathbf{x}; \Delta \mathbf{p})^{-1} \quad (4)$$

from which we then obtain our new $\mathbf{y}^{(\mathbf{p})}$; this entire process is iterated until we obtain convergence for \mathbf{p} . The compositional update is required, as opposed to a simple additive update, because we are solving for the incremental warp update $\mathcal{W}(\mathbf{x}; \Delta \mathbf{p})$ not the parameter update $\Delta \mathbf{p}$. This allows us to pre-compute our Jacobian in Equation 2 at $\mathcal{W}(\mathbf{x}; \mathbf{0})$, rather than at each iteration from $\mathcal{W}(\mathbf{x}; \mathbf{p})$; leading to sizeable computational savings. Please refer to (Baker & Matthews 2001) for more details.

2.1 The Simultaneous Algorithm

An immediate problem one can see with the IC algorithm denoted in Equation 1 is that we make the big assumption that the image $T(\mathcal{W}(\mathbf{x}, \mathbf{0}))$ is a good approximation to $Y(\mathcal{W}(\mathbf{x}, \mathbf{p}^*))$. Obviously, if the object being aligned has considerable intra-class appearance variation (e.g., faces), then this assumption can cause problems. To remedy this situation Baker et al. instead proposed the *simultaneous* algorithm (Baker et al. 2003) which attempts to minimize the following with respect to $\Delta \mathbf{q}$,

$$\|\mathbf{y}^{(\mathbf{p})} - \mathbf{z} - \mathbf{Z}\Delta \mathbf{q}\|^2 \quad (5)$$

where $\Delta \mathbf{q} = [\Delta \mathbf{p}^T, \Delta \lambda^T]^T$ denotes a simultaneous updates in warp $\Delta \mathbf{p}$ and appearance $\Delta \lambda$. We define,

$$\mathbf{z} = \mathbf{t} + \sum_{i=1}^m \lambda_i^{prev} \mathbf{a}_i \quad (6)$$

$$\mathbf{Z}_{\Delta \lambda} = [\mathbf{a}_1, \dots, \mathbf{a}_m] \quad (7)$$

and,

$$\mathbf{Z}_{\Delta \mathbf{p}} = \mathbf{J}(\mathbf{t}) + \sum_{i=1}^m \lambda_i^{prev} \mathbf{J}(\mathbf{a}_i) \quad (8)$$

where \mathbf{a}_i refers to the i th appearance eigenvector, estimated through PCA from an offline ensemble of previously aligned objects (see Figure 1), λ_i^{prev} is the appearance from the previous iteration, and \mathbf{t} denotes

the mean appearance of the offline ensemble. In an analogous result to Equation 3 the solution to Equation 5 is,

$$\Delta \mathbf{q} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y}^{(\mathbf{p})} - \mathbf{z}) \quad (9)$$

where $\mathbf{Z} = [\mathbf{Z}_{\Delta \lambda}, \mathbf{Z}_{\Delta \mathbf{p}}]$. The current warp \mathbf{p} is updated by $\Delta \mathbf{p}$ according to the inverse compositional update in Equation 4 and the current appearance $\lambda = [\lambda_1, \dots, \lambda_m]^T$ is updated by,

$$\lambda = \lambda^{prev} + \Delta \lambda \quad (10)$$

A problem with the simultaneous solution however, is that \mathbf{z} , \mathbf{Z} and therefore $(\mathbf{Z}^T \mathbf{Z})^{-1}$ must be re-estimated at each iteration which slows down the algorithm dramatically. A simple speedup, which Baker et al. refer to as the *project-out* algorithm (Baker et al. 2003), can be found by assuming² that $\lambda^{prev} = \mathbf{0}$ at each iteration which ensures \mathbf{z} and \mathbf{Z} remain constant. In reality since the appearance is not updated at each iteration then $\Delta \lambda$ does not need to be found explicitly.

2.2 The Sequential Algorithm

Although effective, the simultaneous and project-out algorithms suffer some drawbacks due to the simplifying assumptions made in Equations 5-9. The true solution should be to try to solve $\Delta \mathbf{p}$ and λ simultaneously from,

$$\|\mathbf{y}^{(\mathbf{p})} - \mathbf{t} - \mathbf{J}(\mathbf{t}) \Delta \mathbf{p} - \sum_{i=1}^m \lambda_i (\mathbf{a}_i - \mathbf{J}(\mathbf{a}_i) \Delta \mathbf{p})\|^2 \quad (11)$$

Unfortunately, one cannot solve this explicitly for $\Delta \mathbf{p}$ and λ so Baker et al. make the assumption that $\lambda = \Delta \lambda + \lambda^{previous}$ where $\Delta \lambda$ is the appearance update and $\lambda^{previous}$ is the appearance from the previous iteration. Based on this assumption Baker et al. make the approximation,

$$\sum_{i=1}^m \lambda_i \mathbf{J}(\mathbf{a}_i) \approx \sum_{i=1}^m \lambda_i^{prev} \mathbf{J}(\mathbf{a}_i) \quad (12)$$

Thus allowing Equation 11 to be solved simultaneously for $\Delta \mathbf{p}$ and $\Delta \lambda$ instead of for $\Delta \mathbf{p}$ and λ . In this paper we propose a new approach that abandons the approximation in Equation 12 and attempts to solve Equation 11 directly but *not* simultaneously. We refer to this new approach as the *sequential* algorithm. We can pose this algorithm as minimizing with respect to \mathbf{q} the following,

$$\|\mathbf{y}^{(\mathbf{p})} - \mathbf{z}_{\mathbf{q}} - \mathbf{Z}_{\mathbf{q}} \mathbf{q}\|^2 \quad (13)$$

where $\mathbf{q} \in \{\Delta \mathbf{p}, \lambda\}$ as we are attempting to solve for $\Delta \mathbf{p}$ and λ sequentially. Both $\Delta \mathbf{p}$ and λ can be solved in a similar fashion to Equations 3 and 9 where,

$$\mathbf{q} = (\mathbf{Z}_{\mathbf{q}}^T \mathbf{Z}_{\mathbf{q}})^{-1} \mathbf{Z}_{\mathbf{q}}^T (\mathbf{y}^{(\mathbf{p})} - \mathbf{z}_{\mathbf{q}}) \quad (14)$$

First, we attempt to solve for $\Delta \mathbf{p}$ given that we know λ which we initially guess to be $\lambda = \mathbf{0}$,

$$\mathbf{z}_{\Delta \mathbf{p}} = \mathbf{t} + \sum_{i=1}^m \lambda_i \mathbf{a}_i \quad (15)$$

²Please note that the project-out algorithm mentioned here is a slight variation upon the one seen in (Baker et al. 2003), as we are solving for $\Delta \mathbf{p}$ and $\Delta \lambda$ simultaneously rather than sequentially. Empirically however, we have found the performance of these two variants to be identical. Please refer to Appendix A for more details.

$$\mathbf{Z}_{\Delta \mathbf{p}} = \mathbf{J}(\mathbf{t}) + \sum_{i=1}^m \lambda_i \mathbf{J}(\mathbf{a}_i) \quad (16)$$

Given that we have an estimate for $\Delta \mathbf{p}$, from Equation 14, we then obtain a new estimate of $\mathbf{y}^{(\mathbf{p})}$ by applying the inverse compositional warp in Equation 4. We can next solve for λ given our new estimate of $\Delta \mathbf{p}$ where,

$$\mathbf{z}_{\lambda} = \mathbf{t} \quad (17)$$

and,

$$\mathbf{Z}_{\lambda} = [\mathbf{a}_1, \dots, \mathbf{a}_m] \quad (18)$$

The algorithm is iterated until \mathbf{p} and λ reach convergence. As mentioned previously, the algorithm first solves for $\Delta \mathbf{p}$ then solves for λ . The sequential algorithm offers substantial computational savings over the simultaneous algorithm. First, it factorizes the inversion of $(\mathbf{Z}^T \mathbf{Z})$ in Equation 9 into $(\mathbf{Z}_{\Delta \mathbf{p}}^T \mathbf{Z}_{\Delta \mathbf{p}})^{-1}$ and $(\mathbf{Z}_{\lambda}^T \mathbf{Z}_{\lambda})^{-1}$. Second, since \mathbf{Z}_{λ} contains only eigenvectors then $(\mathbf{Z}_{\lambda}^T \mathbf{Z}_{\lambda})^{-1} = \mathbf{I}$ thus making this inversion pointless, again adding considerably to the computational savings over the simultaneous algorithm. Finally, the number of image warps per iteration remains exactly the same as the simultaneous algorithm.

One could argue that their may be some benefit in first estimating λ then estimating $\Delta \mathbf{p}$, in that if the current estimate \mathbf{p} is close to the true alignment \mathbf{p}^* then estimating λ first will allow one to then gain a much more accurate $\Delta \mathbf{p}$. However, we contend this argument is very dependent on the assumption that your current \mathbf{p} is close to \mathbf{p}^* . Empirically we found the more cautious view that \mathbf{p} may be some distance away from \mathbf{p}^* to give more robust results.

3 Frontal-Face Experiments

For our experiments with frontal faces we ran face alignment experiments on the FRGC 1.0 database that corresponded to the training portion of Experiment 1 (Phillips, Flynn, Scruggs, Bowyer, Chang, Hoffman, Marques, Jaesik & Worek 2005). Of the 152 images in this set 76 were used for learning the mean template and appearance variation (i.e. eigenvectors) and the other 76 were used for evaluation. All the images had three hand annotated fiducial points centered on the eyes and nose for ground truth.

3.1 Synthetic Alignment Noise

For our first lot of experiments an initial alignment error was introduced by adding random Gaussian noise to all three points so the *total point error* (TPE) was equal to: (a) 10 pixels and (b) 20 pixels. The TPE is defined as the total distance, in pixels, the current warp's points are from the ground-truth. The TPE is always taken with reference to the template, which was chosen to be of size 80×80 pixels. A comparison between the IC algorithms with: (i) *no appearance variation* (i.e., just the mean template), (ii) *project-out*, (iii) *simultaneous* and (iv) *sequential*, can be seen in Figures 2 and 3.

Figure 2 depicts the ideal scenario where all facial appearance variation has been observed offline previously. Figure 3 depicts the "real world" scenario where the facial appearance variation has not been observed previously. One can see in both figures that the simultaneous and sequential algorithms perform best in all cases. Interestingly, one can see in Figure 2, for the case where the appearance variation

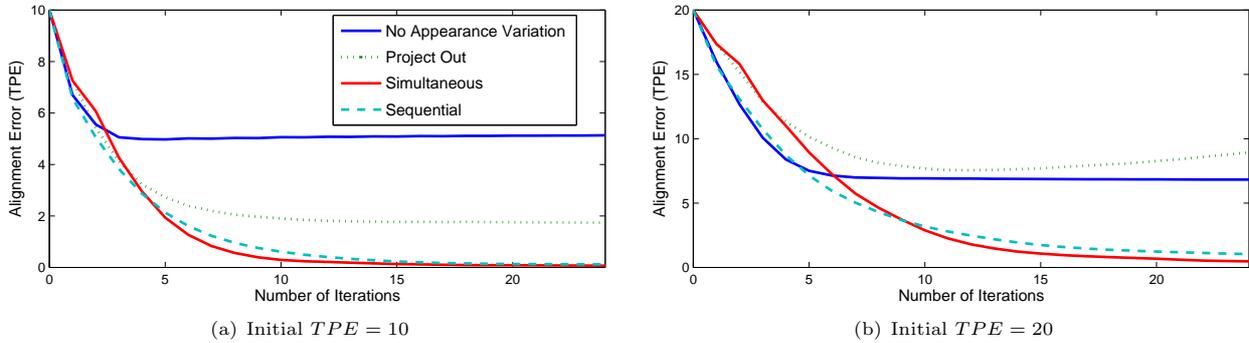


Figure 2: This figure depicts a comparison between IC algorithms when the appearance variation has been seen previously offline. Specifically we compare cases for: (i) *no appearance variation* (i.e. just the mean template), (ii) *project-out*, (iii) *simultaneous* and (iv) *sequential*. Results indicate that the simultaneous and sequential algorithms converge to almost perfect alignment. Our proposed sequential algorithm however, has considerably less computational load than the simultaneous algorithm.

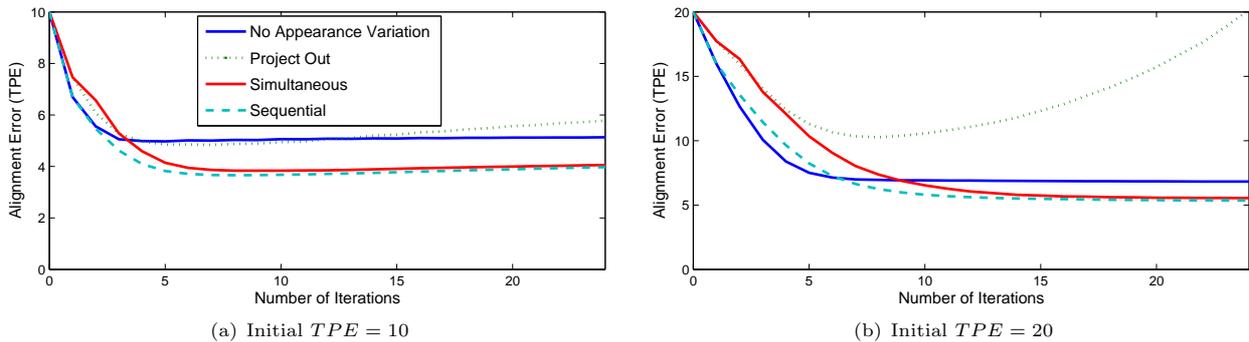


Figure 3: This figure depicts a comparison between IC algorithms when the appearance variation has *not* been seen previously offline. Specifically we compare cases for: (i) *no appearance variation* (i.e. just the mean template), (ii) *project-out*, (iii) *simultaneous* and (iv) *sequential*. Results indicate that although both the simultaneous and sequential algorithms obtain the best alignment, performance is still poor when compared to the results seen in Figure 2 for the scenario when the appearance variation has been seen previously offline.

was seen offline, the final alignment error is almost zero. The biggest contrasts in performance between Figures 2 and 3 can be seen for the project-out algorithm. When the appearance variation has been previously observed, the final alignment error is reasonable. However, for the “unseen” scenario, where the appearance variation has not been observed, the final alignment error diverges. This poor result can be attributed to the assumptions made in the project-out algorithm; namely that the appearance λ^{prev} at each iteration is always zero.

A major result from the experiments carried out in Figures 2 and 3 is the approximately equivalent performance of the sequential and simultaneous algorithms. This result is initially perplexing, as one would expect in most cases a simultaneous iterative solution to be more accurate, since we are solving for appearance and warp at the same time, than a sequential one. This result is consistent for when the initial alignment error is small ($TPE = 10$) and large ($TPE = 20$), as well as for the scenarios where the appearance variation was and was not observed respectively.

3.2 Viola-Jones Noise

In our next lot of experiments we decided to employ an exhaustive search face detector as an initializer, to get an indication of the advantages of our proposed system. The exhaustive search face detector we employed in our experiments was the publicly avail-

able implementation of the Viola-Jones face detector (Viola & Jones 2001) from the OpenCV library. The face detector outputs a bounding box defined by $[x, y, s]$, where $[x, y]$ defines the center of the box and s defines its scale. To gain a good “rough” estimate of where the fiducial points of the face are, based on this bounding box, a projection matrix is learnt that maps from this bounding box to the estimated fiducial face points. This projection matrix is learnt through least-squares optimization from an ensemble of offline aligned face images (see Figure 1).

Figure 4 depicts the distribution, in TPE, of the initial Viola-Jones and also the final distribution after we post-process these coordinates with our gradient descent method. One can clearly see that our method successfully reduces the mean TPE of the Viola-Jones detector. An example of some of these improved alignments can be seen in Figure 5.

4 Non-Frontal Face Experiments

Experiments were performed on a subset of the FERET database (Phillips, Moon, Rizvi & Rauss 2000), specifically images stemming from the *ba*, *bb*, *bc*, *bd*, *be*, *bf*, *bg*, *bh*, and *bi* subsets; which approximately refer to rotations about the vertical axis of 0° , $+60^\circ$, $+40^\circ$, $+25^\circ$, $+15^\circ$, -15° , -25° , -40° , -60° respectively. The database contains 200 subjects in total, which were randomly divided into offline training and online testing sets both



Figure 5: This figure contains examples images for: (a) Viola-Jones alignment, (b) gradient descent alignment, and (c) ground truth alignment. As one can see from these images, our algorithm performs a good job in estimating the correct alignment across a number of different view points.

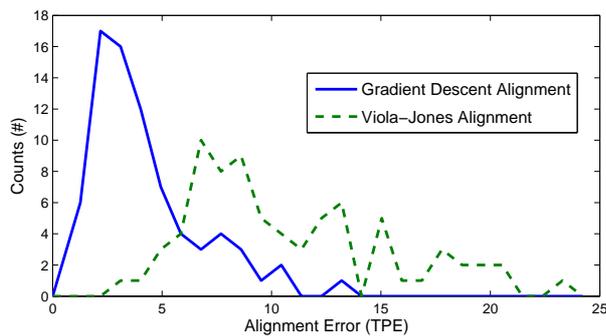


Figure 4: This figure depicts the distribution, in total point error (TPE), of the Viola-Jones face detector alignment and our proposed gradient descent alignment. One can see clearly that our approach both decreases the mean and variance of the TPE produced by the Viola-Jones detector alone.

containing 100 subjects. In a similar fashion to the frontal face experiments templates for all poses were chosen to be of size 80×80 pixels.

In Figure 6 one can see results in terms of the average TPE across a number of different poses. We compare the TPE obtained from the Viola-Jones face detector and our gradient descent method. One can see in all cases our gradient descent method improves the average TPE. Although not perfect, our gradient descent refiner is able to substantially improve face alignment from multiple view-points. Examples of aligned images, from all poses, can be seen in Figure 5 for: (a) the Viola-Jones alignment, (b) our gradient descent alignment, and (c) the ground truth alignment.

5 Conclusion and Future Work

We presented a novel and effective approach to face refinement, on frontal and non-frontal faces, based on a gradient descent image alignment paradigm with appearance variation. Our approach is able to overcome some of the inherent computational difficulties associated with exhaustive search type object detectors when one wants to align an object with more degrees of freedom than just translation and scale. It is also a viable alternative to approaches that rely on affine invariant descriptors (e.g., the eyes) within the object, especially when the location and nature of these descriptors are unclear for the object (e.g., non-

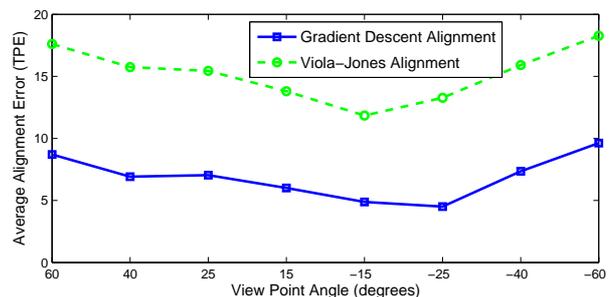


Figure 6: This figure depicts performance in terms of average total point error (TPE) for Viola-Jones alignment and our own gradient descent method across many view-points. One can see that our technique improves the average TPE across all poses.

frontal faces). In this work we proposed an efficient extension to current algorithms in literature, which we refer to as the *sequential* algorithm. This approach was able to empirically deliver approximately the accuracy of the simultaneous algorithm with much less computational cost; making it of viable use in many real-time face processing applications that require human and computer interaction. As a proof of concept we were able to demonstrate how effectively our approach performs in conjunction with a Viola-Jones face detector on frontal and non-frontal faces. We want to extend our current work to deal with more alignment points and more complicated warps (e.g., piece-wise affine) involving faces across pose.

References

- Baker, S., Gross, R. & Matthews, I. (2003), Lucas-kanade 20 years on: A unifying framework: Part 3., Technical Report CMU-RI-TR-03-01, Robotics Institute, Carnegie Mellon University.
- Baker, S. & Matthews, I. (2001), Equivalence and efficiency of image alignment algorithms, in 'IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 1090–1097.
- Black, M. & Jepson, A. (1998), 'Eigen-tracking: Robust matching and tracking of articulated objects using a view-based representation', *International Journal of Computer Vision* **36**(2), 101–130.

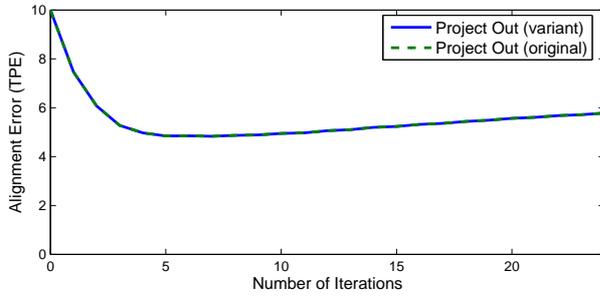
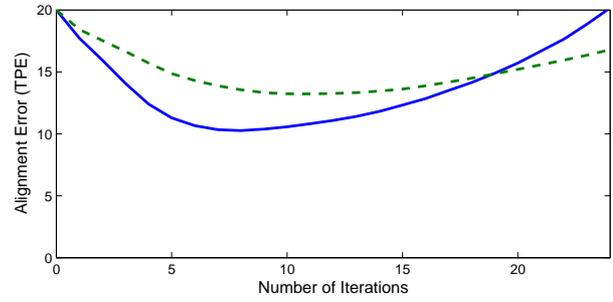
(a) Initial $TPE = 10$ (b) Initial $TPE = 20$

Figure 7: This figure depicts a comparison between two variants of the project-out algorithm, where experiments were conducted on the frontal faces of the FRGC dataset. We note that for a smaller initial alignment error ($TPE = 10$) the performance of our variant and the original is identical (see (a)). However, for a larger initial alignment error ($TPE = 20$) the performance of our variant is superior to the original. Both approaches exhibit poor performance however, as they diverge from the initial alignment.

Everingham, M. & Zisserman, A. (2006), Regression and classification approaches to eye localization in face images, *in* ‘International Conference on Automatic Face and Gesture Recognition’, pp. 441–446.

Gross, R., Baker, S. & Matthews, I. (2005), ‘Generic vs. person specific active appearance models’, *Image and Vision Computing* **23**(11), 1080–1093.

Hager, G. D. & Belhumeur, P. N. (1998), ‘Efficient region tracking with parametric models of geometry and illumination’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(10), 1025.

Lowe, D. G. (1999), Object recognition from local scale-invariant features, *in* ‘IEEE International Conference on Computer Vision’, Vol. 2, pp. 1150–1157.

Lucas, B. & Kanade, T. (1981), An iterative image registration technique with an application to stereo vision, *in* ‘International Joint Conference on Artificial Intelligence’, pp. 674–679.

Moghaddam, B. & Pentland, A. (1997), ‘Probabilistic visual learning for object recognition’, *IEEE Trans. PAMI* **19**(7), 696–710.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Jaesik, M. & Worek, W. (2005), Overview of the face recognition grand challenge, *in* ‘Computer Vision and Pattern Recognition (CVPR)’, pp. 947–954.

Phillips, P. J., Moon, H., Rizvi, S. A. & Rauss, P. J. (2000), ‘The FERET evaluation methodology for face-recognition algorithms’, *IEEE Trans. PAMI* **10**(22), 1090–1104.

Rurainsky, J. & Eisert, P. (2004), Eye center localization using adaptive templates, *in* ‘CVPR Workshop on Face Processing in Video’.

Viola, P. & Jones, M. (2001), Rapid object detection using a boosted cascade of simple features, *in* ‘IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)’, Vol. 1, pp. 511–518.

Wang, P. & Ji, Q. (2005), Learning discriminant features for multi-view face and eye detection,

in ‘IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)’, pp. 373–379.

Appendix

A Variants on Project-Out

In Section 2.1 we present a variant on the project-out algorithm (Baker et al. 2003) first proposed by Baker et al. We proposed in our variant that the formulation of the project-out algorithm can simply be interpreted as the normal simultaneous algorithm, with the exception that we assume λ^{prev} is equal to zero at each iteration. This assumption leads to large computational savings as there is no longer any need for costly matrix inversions at each iteration. This interpretation however, differs slightly to the original formulation of the project-out algorithm. The difference between our formulation and the original project-out algorithm lies in how we minimize,

$$\|\mathbf{y}^{(p)} - \mathbf{t} - \mathbf{J}(\mathbf{t})\Delta\mathbf{p} - \sum_{i=1}^m \Delta\lambda_i \mathbf{a}_i\|^2 \quad (19)$$

with respect to $\Delta\mathbf{p}$ and $\Delta\lambda$. In Baker et al.’s approach they decompose this problem further into the linear subspace $\text{span}(\mathbf{a}_i)$ spanned by the collection of vectors \mathbf{a}_i and its orthogonal complement $\text{span}(\mathbf{a}_i)^\perp$. Baker et al.’s approach is now, with respect to $\Delta\mathbf{p}$ and $\Delta\lambda$, attempting to minimize,

$$\|\mathbf{y}^{(p)} - \mathbf{t} - \sum_{i=1}^m \Delta\lambda_i \mathbf{a}_i\|_{\text{span}(\mathbf{a}_i)}^2 + \|\mathbf{y}^{(p)} - \mathbf{t} - \mathbf{J}(\mathbf{t})\Delta\mathbf{p}\|_{\text{span}(\mathbf{a}_i)^\perp}^2 \quad (20)$$

where $\|\cdot\|_L$ denotes the Euclidean L2 norm of a vector projected into the linear subspace L . Essentially this approach forces the optimization of $\Delta\mathbf{p}$ and $\Delta\lambda$ into two disjoint spaces. One can see that the first term is always exactly zero because the term $\sum_{i=1}^m \Delta\lambda_i \mathbf{a}_i$ can represent any vector in $\text{span}(\mathbf{a}_i)$. As a result the simultaneous minimum over both $\Delta\mathbf{p}$ and $\Delta\lambda$ can be found sequentially by minimizing the second term with respect to $\Delta\mathbf{p}$ alone, and then treating the optimal values of $\Delta\mathbf{p}$ as a constant to minimize the first term with respect to $\Delta\lambda$.

The variant we employed in Section 2.1 actually solves Equation 19 simultaneously for $\Delta\mathbf{p}$ and $\Delta\lambda$

rather than sequentially. Both approaches are extremely fast as nearly all steps can be pre-computed and they require no matrix inversion except in pre-computation. There is a slight computational advantage in Baker et al.'s original formulation as the final update matrix, which one multiplies the error image by, has a rank equal to the dimensionality of just the warp space; whereas our formulation employs an update matrix whose rank is equal to the dimensionality of the warp and appearance space. Empirically we found both approaches obtained identical performance when the initial alignment error is small (see Figure 7(a)), but there is some slight advantage in our approach when the initial alignment error is large (see Figure 7(b)); although in both cases performance did diverge. The experiments in Figure 7 were carried out on the frontal faces of the FRGC dataset.

Learning Active Appearance Models from Image Sequences

Jason Saragih¹Roland Goecke^{1,2*}¹Research School of Information Sciences and Engineering, Australian National University²National ICT Australia, Canberra Research Laboratory
Canberra, Australia

Email: jason.saragih@rsise.anu.edu.au, roland.goecke@nicta.com.au

Abstract

One of the major drawbacks of the Active Appearance Model (AAM) is that it requires a training set of pseudo-dense correspondences. Most methods for automatic correspondence finding involve a groupwise model building process which optimises over all images in the training sequence simultaneously. In this work, we pose the problem of correspondence finding as an adaptive template tracking process. We investigate the utility of this approach on an audio-visual (AV) speech database and show that it can give reasonable results.

Keywords: AAM, automatic model building.

1 Introduction

Active appearance models (AAM) are a powerful class of generative parametric models for non-rigid visual objects which couple a compact representation with an efficient alignment method. Since its advent by Edwards *et al.* in (Edwards, Taylor & Cootes 1998) and their preliminary extension (Cootes, Edwards, Taylor, Burkhardt & Neuman 1998), the method has found applications in many image modelling, alignment and tracking problems, for example (Lehn-Schiöler, Hansen & Larsen 2005) (Stegmann & Larsson 2003) (Mittrapiyanuruk, DeSouza & Kak 2005).

The main drawback of AAM is that it requires pseudo-dense annotations for every training image to build its statistical models of shape and texture. Each of these images may require hundreds of corresponding points. Manual annotation for large databases, therefore, are tedious and error prone. The process is especially difficult for objects which exhibit only a small number of corner like features (i.e. the human face contains mostly edges). A process which automates the annotation process is, hence, highly desirable and may encourage a more widespread utilisation of the AAM.

In this paper, we discuss the automatic annotations (finding physically corresponding points across images) of audio-visual (AV) speech databases which consist of sequences of talking heads. As a test case, we investigate its utility on the AVOZES (Goecke & Millar 2004) database. This scenario for automatic annotations is more constrained than the gen-

eral problem as the changes in shape and texture between consecutive frames in a sequence is relatively small. Nonetheless, we show that this problem is still a challenging one, mainly due to the high dimensionality of the problem which makes it difficult to optimise and avoid spurious local minima.

We approach the automatic annotation process through a tracking perspective, where the annotations in a reference image are propagated through the sequence by virtue of an adaptive template. We begin with an overview of related work in Section 2. The problem of image based correspondences is discussed in Section 3. An outline of our approach to the automatic annotations of image sequences is then presented in Section 4. In Section 5, we describe the results of applying this approach to the AVOZES database. Section 6 concludes with discussions of the results and future directions.

2 Related Work

There has been significant research over the years to automatically find semi-dense correspondences across images of the same class for building AAMs. These methods can be broadly categorised into either feature or image based approaches.

Feature based methods (Chui, Win, Schultz, Duncan & Rangarajan 2003) (Walker, Cootes, & Taylor 1999) (Hill & Taylor 1996) find correspondences between salient features in the image by examining the local structure of the features. The advantage of this method is that feature comparisons and calculations are relatively cheap. The downside however is twofold. Firstly, there may be insufficient salient features in the object to build a good appearance model. Secondly, as the feature comparisons generally consider only local image structure, the global image structure for which the AAM is then modelled is ignored, and hence, the model may be suboptimal.

Image based methods (Cootes, Marsland, Twinning, Smith & Taylor 2004) (Baker, Matthews & Schneider 2004) (Jebara 2003) usually find dense image correspondences by finding a nonlinear warping function which minimises some type of error measure between the intensities of the images. The main advantage of these methods is that the global structure of the image is taken into account, better mimicking the AAM for which the correspondences will be used later. The main drawback of this approach is that to accurately represent the shape variations of the visual object, the warping function will generally need to be parametrised using a large number of parameters (generally as set of landmark points). This results in a very large optimisation problem which is slow to optimise and prone to terminating in local minima.

*National ICT Australia is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council
Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

3 Image Based Correspondence

The heart of image based methods for correspondence consists of finding a warping function between a set of images such that every location in one image is warped to the same physically meaningful (corresponding) location in all other images. However, as there is no true sense of the physical correspondence of un-annotated images, the quality of a set of warping functions is usually quantified by some measure of model *compactness* built from the warped images. Examples of these measures include MDL (Cootes, Twining, Petrovic, Schestowitz & Taylor 2005), specificity/generalisation (Schestowitz, Twining, Petrovic, Cootes, Crum & Taylor 2006) and minimum volume PCA (Jebara 2003).

Apart from the measure of quality there is a large amount of variation of image based correspondence methods at the implementation level. These variations include, but are not limited to, model and warp parametrisation, model fitting methods and the landmark selection process. In this section, we describe the choices we made on these factors for the experiments presented in Section 5. In most cases, we follow the convention of most AAM implementations.

3.1 Linear Appearance Models

Active appearance models assume the visual phenomenon being modelled takes the form of a degenerate Gaussian distribution, where the shape and texture can be modelled by a compact set of linear modes of variation. The texture is generated as follows:

$$t(\mathbf{x}) = \bar{t}(\mathbf{x}) + \sum_{k=1}^{m_t} q_k t_k(\mathbf{x}), \quad (1)$$

where $t(\mathbf{x})$ is the generated model texture at pixel location \mathbf{x} , $\bar{t}(\mathbf{x})$ is the mean texture at that location, $t_k(\mathbf{x})$ is the k^{th} mode of texture variation and q_k is the magnitude of variation in that direction. Similarly, a novel instance of the model's shape can be generated using:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{k=1}^{m_s} p_k \mathbf{s}_k, \quad (2)$$

where $\mathbf{s} = [\mathbf{x}_1; \dots; \mathbf{x}_n]$ is the shape vector of concatenated landmark locations, $\bar{\mathbf{s}}$ is the mean shape, \mathbf{s}_k is the k^{th} mode of shape variation and p_k is the magnitude of variation in that direction.

These models are usually obtained by applying PCA to a set of annotated images, retaining only the m_t and m_s largest modes of variation in shape and texture respectively. The resulting model is a compact representation of a high dimensional visual object by a small set of parameters.

Although these separate models of variation (called independent appearance models) have shown to adequately represent the variations exhibited by many visual objects, they fail to take into account the correlations between shape and texture. In some cases, where there is a strong correlation between shape and texture, failing to take these correlations into account may result in a model capable of generating unrealistic instances of the object class. Furthermore, the resulting model may not be as compact as it could be, if these correlations are considered in the model building process. An example of this is a person-specific AAM. In these cases, it is beneficial to perform a second level of PCA, this time on the concatenation of the shape and texture parameters:

$$\mathbf{a} = \begin{bmatrix} \mathbf{W}_s \mathbf{p} \\ \mathbf{q} \end{bmatrix}, \quad (3)$$

where \mathbf{W}_s is a weighting matrix which normalises the difference in units between shape and texture. A common choice for this matrix is an isotropic diagonal matrix representing the ratio between the total variations of shape and texture in the training set. By applying PCA to a set of these training vectors, a combined appearance model is obtained, for which novel instances can be generated as follows:

$$\mathbf{a} = \sum_i^{m_a} c_k \mathbf{a}_k, \quad (4)$$

where \mathbf{a}_k is the k^{th} mode of combined appearance variation and c_k is the magnitude of variation in that direction. The combined appearance model can be used to generate novel instances of shape and texture directly as follows:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q}_s \mathbf{a} \quad (5)$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{Q}_t \mathbf{a}, \quad (6)$$

where

$$\mathbf{Q}_s = \mathbf{S} \mathbf{W}_s^{-1} \mathbf{A}_s \quad (7)$$

$$\mathbf{Q}_t = \mathbf{T} \mathbf{A}_t \quad (8)$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_t \end{bmatrix} \quad (9)$$

and

$$\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_{m_s}]$$

$$\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_{m_t}]$$

$$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_{m_a}]$$

are matrices of concatenated modes of variations of shape, texture and appearance, respectively. For visual objects exhibiting strong correlations between shape and texture, the resulting combined appearance model is usually more compact than the independent appearance model, exhibiting a smaller number of modes of variation.

3.2 Model Quality

The quality of a model is usually quantified by some measure of compactness. In our work, we follow the method in (Jebara 2003) which estimates compactness of Gaussian distributed models through an approximation of the volume of the variations of the model. The approximation used here is the determinant of the model's covariance matrix, which is equivalent to the sum of the eigenvalues of the model:

$$Q = \sum_i^m \lambda_i \quad (10)$$

In the AAM, variations in pixel values in the image frame are generated from variations in both shape and texture, each of which is modelled by a Gaussian distribution. Therefore, a measure of compactness of an appearance model must take into account the compactness of both models which may disagree with each other. For example, for the same database, a model which exhibits a compact shape distribution may result in a non-compact texture as it needs to accommodate pixel intensity variations which are not accounted for by the shape. On the other hand, if the texture is evaluated in a reference frame (as opposed to the image frame as is done in an MDL formulation (Cootes et al. 2005)), the shape may be chosen

such that the texture is compact at the cost of a non-compact shape distribution. In (Jebara 2003), only the texture compactness is used as a measure of quality, which may result in a non-compact shape distribution which in turn may result in a model which can generate implausible shapes. Although it is easy to have a single measure of model quality through a weighting of the compactness of shape and texture, this weighting is usually chosen heuristically based on the intuition of good results from manual analysis of example models. In this work, we investigate the trends of the shape and texture compactness measures for different settings of the training parameters.

As a final note, in our implementation the sum in Equation (10) is performed over *all* non-zero eigenvalues of the system rather than only the most significant ones. This is because we want to measure the model quality by considering the total amount of variation in the training set. Since the total variation may differ depending on the implementation details, common methods used in PCA such as retaining only a certain percentage of the total variation may not give a discriminative measure as different amounts of variations may be discarded as noise.

3.3 Landmarks and the Warping Function

The shape of an AAM is defined through a set of landmarks which in turn parametrise the warping function used to project the texture from the image to the reference frame.

3.3.1 Landmark Selection Scheme

The choice of these landmarks is crucial to the compactness of the model. As a rule of thumb, for a given number of landmark points, the set which, under the warping function, accounts for the most amount of shape variation within the object class should be chosen. This way, the variation exhibited in the texture model accredited to shape variation is minimised. However, in the problem of automatic model building, parts of the object which exhibit the most variation in shape are not known a-priori. Therefore, a choice must be made regarding the contribution of each location in the image to the variation in texture due to unaccounted variations in shape.

In general, locations with high texture contribute more to the variation in texture due to unaccounted shape variations than do flat regions. Therefore, we propose using a sequential selective process where landmarks are chosen iteratively based on their saliency, measured by the cornerness of that point in a reference image. This method was adopted in (Cootes et al. 2005), where it was demonstrated that using landmarks on strong edges, and ignoring flat regions, gave the best performance as it allowed more control over the boundary regions in the image. Our method differs however in the way the landmarks are chosen. In their approach, the landmarks are initialised on an equally spaced grid, then moved to the closest strong edge. In our work, we sequentially select the most salient pixel location, then zero-out a small region around that point in the saliency image. This process guarantees that the most salient locations are selected, but prevents trivial landmarks (i.e. those which are too close to represent adequate shape variations) from being selected.

Apart from these salient landmarks, we also add a fixed number of border landmarks, equally spaced around the image border, such that the whole image is encoded into the texture model. As the domain of the texture of an AAM is usually defined within the convex hull of the reference shape only, adding these

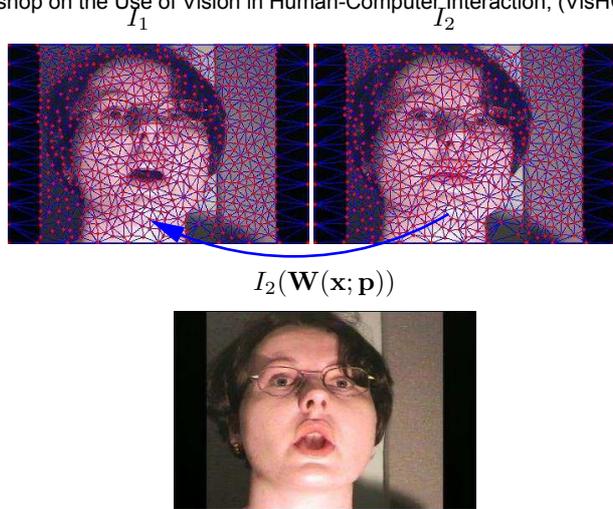


Figure 1: Piecewise-affine warping. Top row: pseudo-dense landmark triangulation. Bottom: I_2 warped onto I_1 using piecewise affine warp defined by triangulation.

border landmarks allow the background to be incorporated into the model's texture which may allow a more accurate model building process as the boundary between the object and the background can give strong cues for model fitting.

3.3.2 Warping Functions

The most common warping function used for AAMs is the piecewise affine warp. This type of warp utilises a triangulation of landmarks in the reference image, where pixels within the domain of each triangle are warped using an affine function. Although there are many other warping function which can be used, such as thin-plate splines or B-Splines, the piecewise affine warp is simple and efficient. Furthermore, it allows the inverse of the warp to be computed efficiently, which is beneficial in an image generation process where the texture in the reference frame is projected onto the image frame.

Although the piecewise affine warp has the disadvantage that it is discontinuous at the boundaries of the triangles, we find that a sufficiently dense set of landmarks chosen according to the scheme described in Section 3.3.1 usually results in a triangulation where the edges in the image correspond to edges of the triangles, minimising the effect of this discontinuity. An example of a pseudo-dense landmark selection with its triangulation and warping process is shown in Figure 1.

3.4 Alignment

Regardless of the model building process used, automatic AAM construction generally involves a non-rigid registration to align the model to an image. The alignment process essentially finds the model parameters which best describe the image. This process usually involves minimising some measure of fitness between the model and the image which contains a data term and a smoothness term:

$$C = C_d + \eta C_s, \quad (11)$$

where C_d is the data term, C_s is the smoothness term and η is a regularisation parameter which trades off the contribution of the data and smoothness terms to the total cost.

3.4.1 The Data Term

The data term is usually defined as a function of the difference between the model's texture and the image texture warped back to the reference frame:

$$C_d = \sum_{\mathbf{x} \in \Omega} \rho(E(\mathbf{x}); \sigma) \quad (12)$$

$$E(\mathbf{x}) = t(\mathbf{x}; \mathbf{q}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})), \quad (13)$$

where Ω is the domain over which the model's texture is defined (i.e. the convex hull of the landmark points), $t(\mathbf{x}; \mathbf{q})$ is the model's texture, $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ is the image texture warped back to the reference frame, and ρ is some type of function over the residuals, parametrised by σ .

A common function used in AAM alignment is the L2-norm (Baker & Matthews 2002), in which case, the data term takes the least squares form. However, in some cases it may be beneficial to use a robust error function to minimise the effects outliers in the data. This is particularly important in model building as regions which are not yet accounted for by the texture model may deteriorate the estimate of the shape in other parts of the image, leading to a non-compact model. For the experiments presented in Section 5, we use the *Geman-McClure* function:

$$\rho(r; \sigma) = \frac{r^2}{\sigma^2 + r^2}, \quad (14)$$

which has been used extensively for optical flow estimation (Black & Anandan 1993) (Blake, Isard & Reynard 1994).

The choice of the scale parameter σ for robust error functions is always problematic as it depends on the distribution of the residuals. One approach is to use the assumption that the corresponding error functions model the underlying distribution of residuals, and find σ which best fits that distribution. However, this usually leads to a complex non-linear estimation process. Therefore, in our work, we assume a contaminated Gaussian distribution for the residuals. In this framework, the estimate of σ can be derived from the median value of the absolute residuals:

$$\sigma = 1.4826 \text{ med}(E(\mathbf{x})) \quad (15)$$

which has been claimed to have excellent resistance towards outliers, tolerating almost 50% of them (Sawhney & Ayer 1996).

3.4.2 The Smoothness Term

In automatic model building the landmarks should be allowed to move freely to minimise the data term. However, as the AAM's shape consists of a pseudo-dense set of landmarks, the dimensionality of the optimisation process is very large, which if not constrained is likely to get trapped in spurious local minima. These minima usually correspond to implausible shapes. As such, a smoothness term is required to encourage the model to deform smoothly.

The form of the smoothness constraint is dependent on the visual object being modelled. The most common of which is to penalise the magnitude of the deformation of every landmark from a reference shape, as was adopted in (Baker et al. 2004). The problem with this approach is that it does not take into account the spatial relationship between the deformation of landmarks. In this work, we penalise only the *difference* between the deformation of landmarks, similar to the smoothness constraint in variational optical flow estimation (Brox, Bruhn, Papenber & Weickert 2004). The differences are weighted

by a smooth function of the landmark distances in a predefined shape:

$$C_s = \sum_{i,j} k_{ij} \|\mathbf{d}(i, j)\|^2, \quad (16)$$

where

$$k_{ij} = \frac{\exp\left(-\frac{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|^2}{2\sigma_s^2}\right)}{\sum_j \exp\left(-\frac{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|^2}{2\sigma_s^2}\right)} \quad (17)$$

is a smoothing factor and

$$\mathbf{d}(i, j) = [\mathbf{W}(\mathbf{x}_i; \mathbf{p}) - \hat{\mathbf{x}}_i] - [\mathbf{W}(\mathbf{x}_j; \mathbf{p}) - \hat{\mathbf{x}}_j] \quad (18)$$

is the difference between landmark displacements, with $\hat{\mathbf{x}}_k = \mathbf{W}(\mathbf{x}_k; \mathbf{p}_0)$ being the location of the k^{th} landmark in the predefined shape, parametrised by \mathbf{p}_0 . In most works utilising a smoothness measure, the predefined shape is always set to the reference shape (i.e. $\mathbf{p}_0 = \mathbf{0}$). The problem with this is that it assumes the deformations are isotropic for all landmarks. This type of smoothing does not fit the notion of a linear shape class which is modelled by a degenerate Gaussian. In contrast, we set the predefined shape as the initial shape in the alignment process. Smoothing the deformations in an isotropic manner starting from this shape better suits the form of the shape model as it does not over constrain the overall highly anisotropic shape deformations whilst still encouraging the landmarks to deform smoothly.

3.4.3 Optimisation

To optimise the cost function in Equation (11) we adopt the Gauss-Newton method which is commonly used for image alignment. To allow the use of the robust error function in the Gauss-Newton optimisation procedure, the data term must be reformulated. Since it contains no *squared* term, the derivation of the parameter update requires a second order Taylor expansion, akin to the Newton algorithm. Therefore, following (Baker, Gross & Matthews 2003), we replace the data term in Equation (12) with:

$$C_d = \sum_{\mathbf{x} \in \Omega} \varrho(E(\mathbf{x})^2; \sigma) \quad (19)$$

and the reformulated robust error function:

$$\varrho(r; \sigma) = \frac{r}{\sigma^2 + r} \quad (20)$$

This requires only that the error function is symmetric, which is satisfied by the Geman McClure function.

With this reformulation, the Gauss-Newton Hessian of the data term is given by:

$$\mathbf{H}_d = \sum_{\mathbf{x} \in \Omega} \varrho'(E(\mathbf{x})^2) \mathbf{J}_d(\mathbf{x})^T \mathbf{J}_d(\mathbf{x}) \quad (21)$$

where $\varrho'(E(\mathbf{x})^2)$ is the derivative of the reformulated robust error function and

$$\mathbf{J}_d(\mathbf{x}) = \left[-\nabla I(\mathbf{W}(\mathbf{x}; \mathbf{p})) \frac{\partial \mathbf{W}(\mathbf{x}; \mathbf{p})}{\partial \mathbf{p}}, \frac{\partial t(\mathbf{x}; \mathbf{q})}{\partial \mathbf{q}} \right] \quad (22)$$

is the Jacobian of the data term. It should be noted here that since we allow the landmark points to move freely, the warping function \mathbf{W} is directly parametrised by the location of the landmarks (i.e.

$\mathbf{p} = [\mathbf{x}_1; \dots; \mathbf{x}_n]$. Therefore, the distance measure in Equation (18) is equivalent to:

$$\mathbf{d}(i, j) = (\mathbf{x}_i - \hat{\mathbf{x}}_i) - (\mathbf{x}_j - \hat{\mathbf{x}}_j) \quad (23)$$

This is in contrast to the usual AAM formulation where the warp is parametrised by the magnitudes of the modes of shape variation.

The Gauss-Newton Hessian of the smoothness term is given by:

$$\mathbf{H}_s = \sum_{i,j}^n k_{ij} [\mathbf{J}_x(i, j)^T \mathbf{J}_x(i, j) + \mathbf{J}_y(i, j)^T \mathbf{J}_y(i, j)] \quad (24)$$

where the $2k^{th}$ entry of the x smoothness term's Jacobian $\mathbf{J}_x(i, j)$ is given by:

$$\mathbf{J}_x(i, j)^{2k} = \begin{cases} 1 & \text{if } k = i \\ -1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}, \quad (25)$$

and similarly for the $(2k + 1)^{th}$ entry of \mathbf{J}_y . For \mathbf{J}_x , entries at the $(2k + 1)^{th}$ locations are all zero, and similarly for the $2k^{th}$ locations of \mathbf{J}_y . This simple form, which affords a fast calculation of the Hessian and gradient, is a result of optimising directly over the landmark locations.

The parameter updates of the Gauss-Newton optimisation of Equation (11) then takes the following form:

$$\begin{bmatrix} \Delta \mathbf{p} \\ \Delta \mathbf{q} \end{bmatrix} = - \left[\mathbf{H}_d + \eta \begin{bmatrix} \mathbf{H}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right]^{-1} \left[\mathbf{g}_d + \eta \begin{bmatrix} \mathbf{g}_s \\ \mathbf{0} \end{bmatrix} \right] \quad (26)$$

where

$$\mathbf{g}_d = \sum_{\mathbf{x} \in \Omega} \rho'(E(\mathbf{x})^2) \mathbf{J}_d(\mathbf{x})^T E(\mathbf{x}) \quad (27)$$

$$\mathbf{g}_s = \sum_{i,j}^n k_{ij} [\mathbf{J}_x(i, j)^T \mathbf{d}_x(i, j) + \mathbf{J}_y(i, j)^T \mathbf{d}_y(i, j)] \quad (28)$$

are the gradients of the data and smoothness term respectively.

The optimisation process can usually be sped-up by using the inverse compositional formulation (Matthews & Baker 2003). By reversing the roles of the model and the image in the data term, the gradients of the data term can be precomputed and hence a large proportion of computation needs to be done only once. The extensions of the inverse compositional image alignment (ICIA) algorithm to robust error norms was proposed in (Baker et al. 2003). With this formulation, the Hessian of the data term cannot be precomputed, despite the fixed gradients, as the derivative of the robust error terms cannot be precomputed. Although an efficient approximation has been derived by assuming spatial coherence of the outliers, this implementation is not particularly effective for automatic model building from databases as the images are generally occlusion free, with outliers stemming mainly from misalignment, image noise, changes in texture not yet accounted for by the texture model, and interlacing effects. The presence of the smoothness term means that the Hessian needs to be updated and inverted at every iteration which is the most costly part of the optimisation when there is a large number of landmark points. Furthermore, for the methods described in Section 4, the model is updated after every image, requiring the gradients to

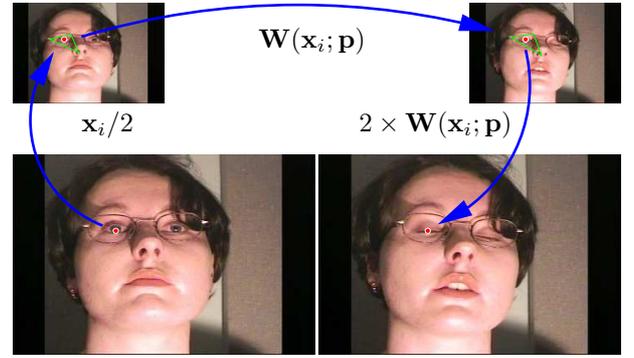


Figure 2: Initialising points in lower levels of the Gaussian pyramid. Top row: Warps at higher pyramid level. Bottom: Landmarks at current pyramid level.

be recomputed. Due to these factors, we predict that using the ICIA formulation will not give dramatic improvements in efficiency and hence it was not incorporated into our implementation.

3.4.4 Gaussian Pyramid

Despite the use of the smoothing term, the optimisation process may still converge to a local minimum due to the high dimensionality of the problem. This problem can be partially alleviated by optimising on a Gaussian-pyramid.

There are issues however with regards to how the shape is parametrised between the levels of the pyramid. A pseudo-dense correspondence at the lowest level of the pyramid may result in an over parametrised model at the highest level of the pyramid, which results both in a slow alignment process as well as the higher likelihood of getting stuck in local minima. Instead, in this work we build a separate model for each level. Starting at the highest pyramid level, a set of landmarks is chosen as described in Section 3.3.1. With this, an automatic model building process described in Section 4 is performed. Moving down the pyramid, a new set of landmarks is chosen from the reference image.

The propagation of these landmarks to other images is illustrated in Figure 2. First the landmarks are downscaled to the previous pyramid level (bottom to top left in Figure 2). Then the landmarks are warped using the found correspondence for that level (top row), and finally up-scaled back to the current pyramid level (top to bottom right).

With the smoothness term described in Section 3.4.2, the use of the Gaussian pyramid allows a stiff regularisation parameter η to be used as the movements of points at every level will be relatively small. This in turn allows the optimisation process to better avoid local minima.

4 Incremental Model Building

Most approaches to automatic model building can be classed as groupwise, where a model is iteratively refined from an initial estimate by first fitting it to each image, followed by a reconstruction of a new model from the fitted images. One of the drawbacks of this approach is that it does not take into account the sequential nature of images in video. As such, its initial estimate of the model may be far from the optimum, which may cause the algorithm to converge slowly or get stuck in local minima.

By assuming that the appearance of the visual objects varies slowly between consecutive frames in a

sequence, the model building process can be posed as a tracking problem. Although the complexity of the warping function is much higher than most tracking problems, which generally solve only for a similarity or affine transform, the same mechanisms apply. We start with an initial template, without loss of generality taken as the first image in the sequence, and propagate the landmark positions to the other images in the sequence through a consecutive alignment process. Unlike typical tracking problems however, due to the high dimensionality of the parameter space, the alignment process must generally utilise gradient based approaches as non-gradient methods such as a particle-filters will be too computationally expensive to evaluate.

One of the main difficulties associated with template tracking is due to the changes in the object's texture throughout the sequence. Although this problem can be partially alleviated by using a robust error function, as the sequence progresses the object's texture may undergo significant changes such that treating them as outliers may lead to misalignment. One solution to this problem is to update the template using the texture from the previous frame. However, simply replacing the texture with the most recent image makes the algorithm prone to drifting. In this work, we investigate the utility of two adaptable template approaches for automatic model building from image sequences.

4.1 Method 1: Grounded Templates

There are a number of approaches to the template update problem which reduce the drifting phenomenon, for example (Matthews, Ishikawa & Baker 2004) (Zhong, Jain & Dubuisson-Jolly 2000) (Loy, Goecke, Rougeaux & Zelinsky 2000). In this work we follow the approach of (Loy et al. 2000), where the new template is defined as a weighted combination of the initial template and the texture from the most recent image:

$$T_t(\mathbf{x}) = \alpha T_0(\mathbf{x}) + (1 - \alpha) T_{t-1}(\mathbf{x}) \quad (29)$$

The parameter $\alpha \in (0 \dots 1)$ is a grounding factor which reduces the drifting effect whilst allowing the template to adapt to the current object's texture.

As the template is updated once before the alignment process in the next image, the optimisation process needs only be done over the landmark locations. Therefore, the Jacobian of the data term in Equation (22) for this method is simply:

$$\mathbf{J}_d(\mathbf{x}) = -\nabla I(\mathbf{W}(\mathbf{x}; \mathbf{p})) \frac{\partial \mathbf{W}(\mathbf{x}; \mathbf{p})}{\partial \mathbf{p}} \quad (30)$$

and the Gauss-Newton update in Equation (26) is now given by:

$$\Delta \mathbf{p} = -[\mathbf{H}_d + \eta \mathbf{H}_s]^{-1} [\mathbf{g}_d + \eta \mathbf{g}_s] \quad (31)$$

The output of the template matching algorithm is a set of corresponding annotations in every image in the sequence, from which an appearance model can be built in the usual manner.

4.2 Method 2: Incremental Texture Learning

One of the weaknesses of the template update approach is that it takes into account only the initial and most recently encountered textures. As such it makes no use of the knowledge of the variations in texture which have been encountered earlier in the sequence. One possibility to incorporate this information is to

perform an incremental model building process as the object is tracked throughout the sequence.

For this algorithm we utilise incremental PCA (Li 2004) to update the model, rather than the template, after matching to every new image. Starting with the template of the first image, we match it to the next image using the approach described in Section 4.1. Some of the variations captured as outliers may in fact be intrinsic variations of the object rather than just image noise. The texture of the newly aligned image is then used as a new data instance for the linear model, for which incremental PCA is used to integrate it into the model. The amnesic factor (a weighting between the current model and the new data instance) is set to $\frac{n}{1+n}$, where n is the number of samples used to build the current model, so that every sample integrated into the model is given the same importance. See (Li 2004) for details.

Once the model exhibits some linear modes of variation apart from the mean, matching to the next image should be done by simultaneously updating the landmark locations and the texture parameters \mathbf{q} using the update equations described in Section 3.4.3. This way, images which exhibit texture variations previously encountered in the sequence will be matched better than using a fixed template. Again, the data term is formulated using the robust error function to account for texture variations not yet encountered in the sequence.

5 Experiments

5.1 The AVOZES Database

AVOZES (Goecke & Millar 2004), the Audio-Video Australian English Speech data corpus, is a database of 20 speakers uttering a variety of phrases which was designed for research on the statistical relationship of audio and video speech parameters with an audio-video automatic speech recognition task in mind. Although sparse annotations for the vital mouth points, such as lip corners, are available, these points are chosen manually and represent only a heuristic intuition about their usefulness for automatic speech recognition. A more elaborate set of cues may be useful for audio-video speech recognition which may not be directly obvious. AAMs, which encode both a pseudo-dense set of landmark points as well as texture variations, provide a rich set of features to a speech recognition system which may allow better recognition rates to be achieved. An intensive study of the application of AAMs in this domain can be found in (Neti, Potaminos, Luetin, Matthews, Glotin & Vergyri 2001).

In our experiments we used the continuous speech sequences for each of the speakers exclusively. The continuous speech part of AVOZES consists of three sequences, each with a different phrase. The length of all sequences range from 90 to 150 frames. As the video files in the database consist of a stereo pair, warped to half the height, we used only part of the sequence pertaining to the left camera, which we scaled to the true ratio.

For each of the speakers, we performed both of the image based correspondence methods described in Section 4 on all three sequences together. Since there may be large differences between the start and end of different sequences of the different phrases, we find the image which is most similar in the later two sequences to an image in the first sequence. After tracking through the first sequence the model is tracked in the other sequence starting from the most similar image found previously, initialising the shape estimate to the corresponding image in the first sequence. The

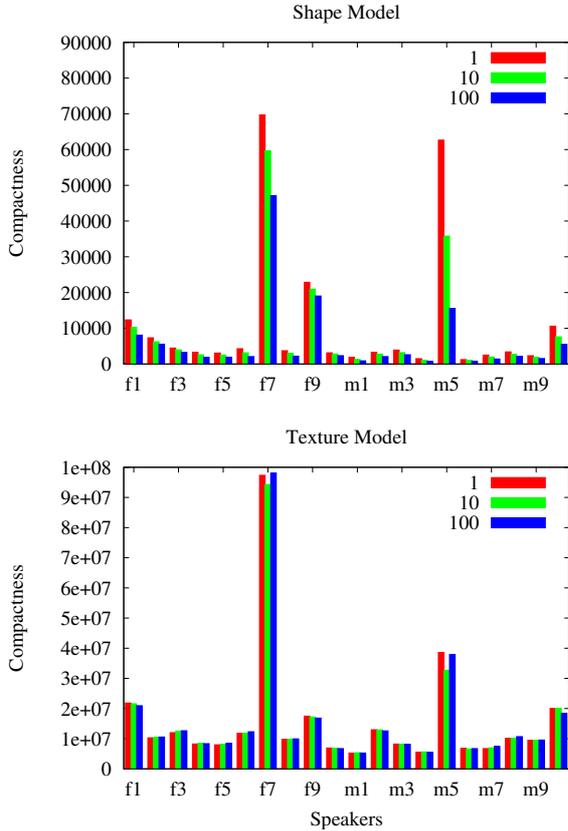


Figure 3: Shape and texture model compactness for every speaker in AVOZES. The models were built from correspondences found using the grounded template method with three settings of the regularisation parameter $\eta = \{1, 10, 100\}$.

tracking process in these other sequences is performed forwards and backwards until the beginning and end of the sequences respectively. From the resulting correspondences, the compactness of the shape and texture models are calculated as described in Section 3.2. The experiments were repeated for a number of settings of the smoothing parameter η .

5.2 Results

In Figure 3 and 4, histograms of the shape and texture model compactness of each of the speakers in the AVOZES database built from correspondences obtained using the methods described in Section 4.1 are shown for three different settings of the regularisation parameter η . Comparing the two methods, the shape compactness differs little between them. The main difference lies in the texture compactness, where the incremental texture learning method generates models which are around twice as compact for most speakers compared to the grounded template method. As discussed in Section 4.2, this result is expected as the incremental texture learning retains memory of previously encountered texture variations. Also, as the alignment process may contain errors which may accumulate throughout the sequence, this approach is more constrained to valid texture instances rather than just the first and most recently encountered texture, which may be erroneous.

Studying each method independently, as expected the compactness of the shape model improves as η is increased. Perhaps somewhat more surprisingly, the texture model's compactness is effected little by the different settings of the regularisation parameter. We attribute this to the fact that the texture model is

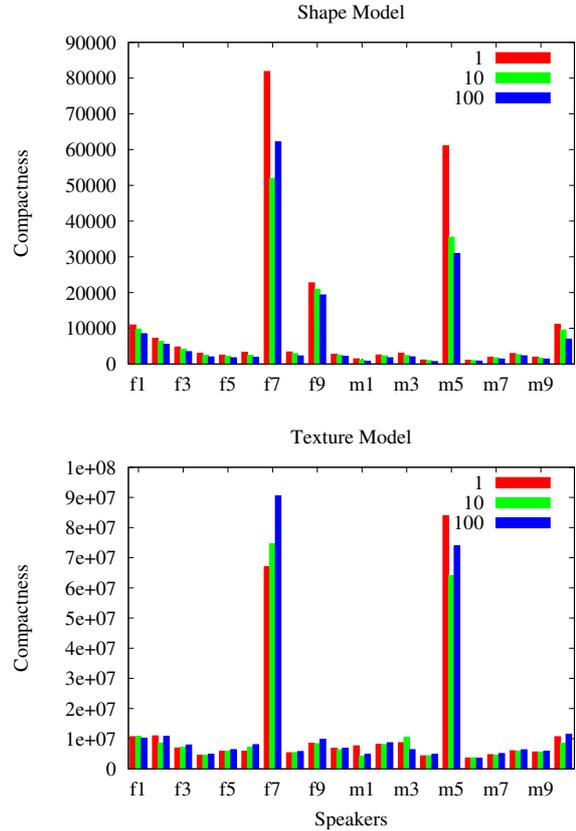


Figure 4: Shape and texture model compactness for every speaker in AVOZES. The models were built from correspondences found using the incremental texture learning method with three settings of the regularisation parameter $\eta = \{1, 10, 100\}$.

evaluated in a reference frame. The effect of this is that for groups of landmarks which correspond to *flat* parts of the image, their movements contribute little to the change in the texture when projected onto the reference frame. As such, shapes with significantly different landmark locations in these flat regions may result in very similar texture. An example of this is shown in Figure 5. Landmarks in flat regions are more likely to be perturbed by image noise and hence, for the same texture compactness, the model with better shape compactness is generally a better model.

From the correspondences in each image, found using the incremental texture learning method with $\eta = 100$, we built a combined appearance model (see Section 3.1) using every 10^{th} image in the sequences. The mean and first mode of variation on all speakers are shown in Figure 7 and 8. Although the correspondences appear to be of high quality in most speakers, observed through the *crispness* of the images, there are a few for which the tracking method seemed to have failed to obtain the correct correspondences across the sequences. In particular, the f7 and m5 speakers are particularly poor, where the first mode of variation seems to entail the presence or disappearance of visual artefacts. Referring to the texture compactness histograms in Figure 3 and 4 it can be seen that these two speakers exhibit the least compact model out of the database by a significant margin.

It is clear that in these cases, the tracking process used to find the correspondences failed significantly in parts of the image, resulting in the texture model needing to account for variations due to misalignment rather than intrinsic texture variations of

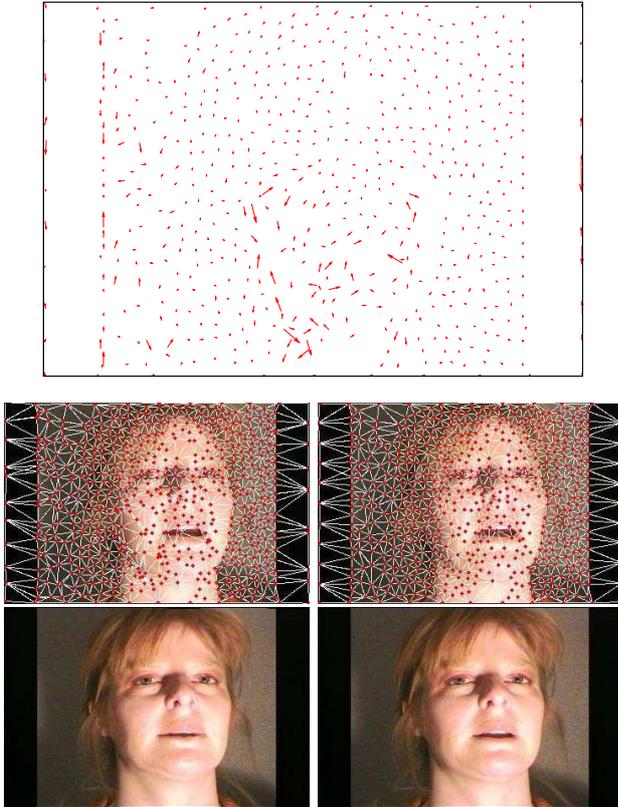


Figure 5: Two shapes with significant landmark differences in *flat* regions exhibiting similar texture when projected to the reference frame. Top: Shape difference. Middle: Shapes of two images. Bottom: Texture projections onto the reference image.

the speaker. On closer inspection, we found that these two speakers exhibited significant motion towards the camera during some parts of the sequences. Example images from these sequences are shown in Figure 6. As such, significant parts of the background are occluded when the speakers are close to the camera, but reappear when they are further from it. Because the background exhibits some strong texture and colour variations (see the white strip behind the speaker's heads), the disappearance/emergence of these areas perturb the alignment process significantly, despite using a robust error function.

As models of the other speakers, which exhibit relatively small amounts of head movement, were able to be built compactly, we suspect that databases which exhibit a uniform background do not exhibit this problem. However, in cases where this is not practical, one solution would be to initialise the feature points within the face region exclusively, either using a manual crop in the first image or using some type of skin colour detector. It should be noted however, that the accuracy of the alignment around the peripheral of the face using this approach may be inferior to that which encodes background.

As a final note, although the methods tested here have shown to give reasonably compact models when no significant visual artifacts disappear or emerge throughout the sequence, because the correspondences are obtained in a pairwise manner the model quality may be improved through a groupwise method. In fact, the methods discussed in this work can be used as a good initialisation for groupwise methods which will encourage faster convergence and help avoid local minima.



Figure 6: Images from the f7 and m5 speakers which illustrate the large differences in scale affecting content in the images.

6 Conclusion

In this work, we have investigated the utility of adaptive tracking methods for automatically building pseudo-dense correspondences across a sequence of a deformable object, with an AV database as a test case. We compared two methods, the grounded template and incremental texture learning method, measuring their performance through a shape and texture compactness measure as well as a qualitative analysis of the resulting linear models of variation.

Through extensive experiments we have shown that this approach can be used to build highly compact models of a linearly deforming object which includes the background in the image. We also found that if the background exhibits significant texture, despite being static, movements of the object which causes these textured regions to be occluded or new textured regions to appear later in the sequence, significantly degrades the performance of this method. However, we suspect that this is a problem exhibited by most image based correspondence methods which utilise diffeomorphic warps and do not explicitly model the disappearance or emergence of visual artifacts.

Future work on extending this method might involve investigations into efficiency gains of using the inverse compositional formulation, evaluating alignment error in the image rather than reference frame, and extensions to incremental shape model learning. Although the methods investigated in this paper and their possible extensions allow significant savings on human intervention, requiring only one manual markup per speaker, for large databases containing thousands of sequences this approach may be infeasible. The much more difficult problem of finding correspondences across sequences of different instances of the same object class (different speakers in AVOZES, for example) remains an open problem.

References

- Baker, S., Gross, R. & Matthews, I. (2003), Lucas-Kanade 20 years on: A unifying framework: Part 3, Technical report, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.
- Baker, S. & Matthews, I. (2002), Lucas-kanade 20 years on: A unifying framework: Part 1, Technical Report CMU-RI-TR-02-16, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

- Baker, S., Matthews, I. & Schneider, J. (2004), 'Automatic construction of active appearance models as an image coding problem', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(10), 1380–1384.
- Black, M. & Anandan, P. (1993), The robust estimation of multiple motions: Affine and piecewise-smooth flow fields, Technical report, Xerox PARC.
- Blake, A., Isard, M. & Reynard, D. (1994), Learning to track curves in motion, in 'IEEE Conference on Decision Theory and Control', pp. 3788–3793.
- Brox, T., Bruhn, A., Papenbergh, N. & Weickert, J. (2004), High accuracy optical flow estimation based on theory of warping, in T. Pajdla & J. Matas, eds, '8th European Conference on Computer Vision', Vol. 4, Springer-Verlag, Prague, Czech Republic, pp. 25–36.
- Chui, H., Win, L., Schultz, R., Duncan, J. S. & Rangarajan, A. (2003), 'A unified non-rigid feature registration method for brain mapping', *Medical Image Analysis* **7**(2), 113–130.
- Cootes, T. F., Edwards, G., Taylor, C. J., Burkhardt, H. & Neuman, B. (1998), Active appearance models, in 'European Conference on Computer Vision', Vol. 2, pp. 484–489.
- Cootes, T. F., Marsland, S., Twining, C. J., Smith, K. & Taylor, C. J. (2004), Groupwise diffeomorphic non-rigid registration for automatic model building, in 'European Conference on Computer Vision', pp. 316–327.
- Cootes, T. F., Twining, C. J., Petrovic, V., Schestowitz, R. & Taylor, C. J. (2005), Groupwise construction of appearance models using piece-wise affine deformations, in 'British Machine Vision Conference', Vol. 2, pp. 879–888.
- Edwards, G., Taylor, C. J. & Cootes, T. F. (1998), Interpreting face images using active appearance models, in 'IEEE International Conference on Automatic Face and Gesture Recognition', pp. 300–305.
- Goecke, R. & Millar, J. B. (2004), The audio-video australian english speech data corpus avozes, in '8th International Conference on Spoken Language Processing INTERSPEECH 2004 - IC-SLP', Vol. III, ISCA, Jeju, Korea, pp. 2525–2528.
- Hill, A. & Taylor, C. J. (1996), A method of non-rigid correspondence for automatic landmark identification, in '7th British Machine Vision Conference', Vol. 2, pp. 323–332.
- Jebara, T. (2003), Images as bags of pixels, in 'International Conference on Computer Vision', pp. 265–272.
- Lehn-Schiøler, T., Hansen, L. K. & Larsen, J. (2005), Mapping from speech to images using continuous state space models, in 'Lecture Notes in Computer Science', Vol. 3361, Springer, pp. 136 – 145.
- Li, Y. (2004), 'On incremental and robust subspace learning', *Pattern Recognition* **37**(7), 1509–1518.
- Loy, G., Goecke, R., Rougeaux, S. & Zelinsky, A. (2000), Stereo 3D lip tracking, in '6th International Conference on Control, Automation, Robotics and Vision', Singapore.
- Matthews, I. & Baker, S. (2003), Active appearance models revisited, Technical Report CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.
- Matthews, I., Ishikawa, T. & Baker, S. (2004), 'The template update problem.', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 810–815.
- Mittrapiyanuruk, P., DeSouza, G. N. & Kak, A. C. (2005), Accurate 3D tracking of rigid objects with occlusion using active appearance models, in 'IEEE Workshop on Motion and Video Computing', pp. 90–95.
- Neti, C., Potamios, G., Luetin, J., Matthews, I., Glotin, H. & Vergyri, D. (2001), Large-vocabulary audio-visual speech recognition: A summary of the john hopkins summer 2000 workshop, in 'Workshop on Multimedia Signal Processing (MMSP)', Cannes.
- Sawhney, H. S. & Ayer, S. (1996), 'Compact representation of videos through dominant and multiple motion estimation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(8), 814–830.
- Schestowitz, R. S., Twining, C. J., Petrovic, V. S., Cootes, T., Crum, B. & Taylor, C. J. (2006), Non-rigid registration assessment without ground truth, in 'Medical Image Understanding and Analysis', Vol. 2, pp. 151–155.
- Stegmann, M. B. & Larsson, H. B. (2003), Fast registration of cardiac perfusion MRI, in 'International Society of Magnetic Resonance In Medicine', Toronto, Canada, p. 702.
- Walker, K. N., Cootes, T. F., & Taylor, C. J. (1999), Automatically building appearance models from image sequences using salient features, in D. T.Pridmore, ed., 'British Machine Vision Conference', Vol. 2, pp. 463–562.
- Zhong, Y., Jain, A. K. & Dubuisson-Jolly, M. P. (2000), 'Object tracking using deformable templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(5), 544–549.

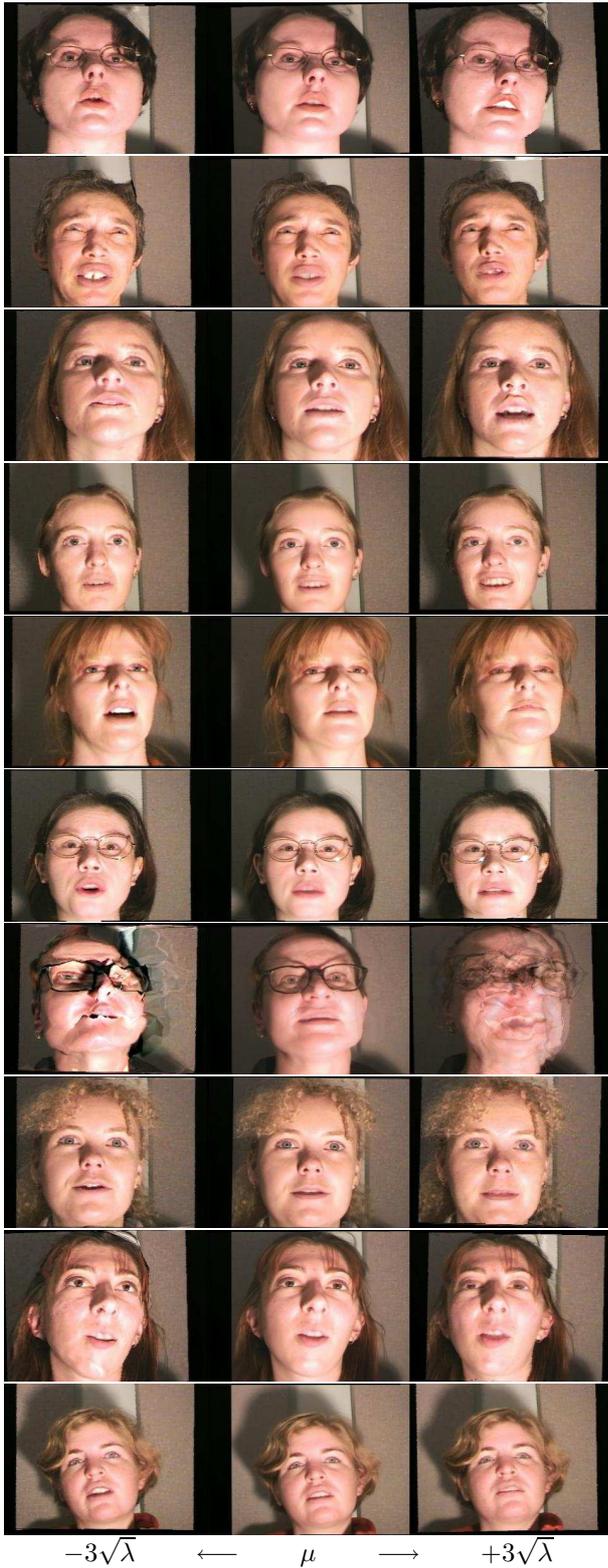


Figure 7: The first mode of variation of the female speakers in AVOZES, varied between \pm three standard deviations.

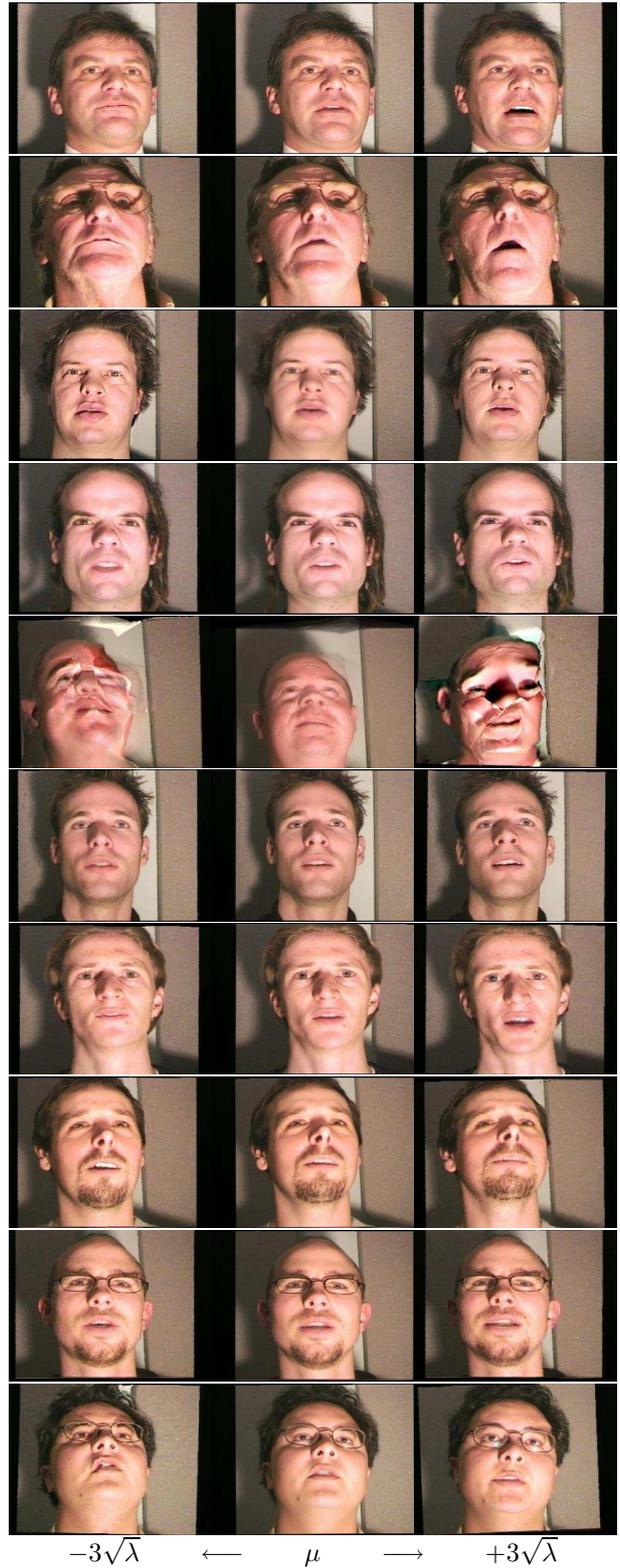


Figure 8: The first mode of variation of the male speakers in AVOZES, varied between \pm three standard deviations.

Using Optical Flow for Step Size Initialisation in Hand Tracking by Stochastic Optimisation

Desmond Chik^{1,2}

¹National ICT Australia, Canberra Research Laboratory*

²Research School of Information Sciences and Engineering, Australian National University
Canberra, Australia

Email: desmond.chik@rsise.anu.edu.au

Abstract

In this paper we use optical flow to initialise step sizes in a stochastic optimisation setting for hand tracking. One can appreciate that most complex hand gestures result in different segments of the hand moving at different speeds. We show that by reflecting this motion difference in our step size initialisation process, the tracking performance can be improved. Significant improvement to tracking accuracy has been observed for gripping motions of the hand.

Keywords: Hand tracking, SMD, Step size initialisation.

1 Introduction

Research into the tracking of articulated structures, such as the human body or the hand, has traditionally been spurred on by interests from the movie animation industry and the sports analysis community. However, with the advent of faster and cheaper computers, the prospect of an inexpensive motion capture system is becoming a realistic possibility for the average consumer. Consequently this type of technology has direct applications to the field of human computer interaction (HCI) as a new medium with which users can interact with computers, such as through natural gestural movements. Sign language recognition for example is an aspect of HCI where hand tracking is applicable to (Holden *et al.* 2005).

Approaches to the tracking of articulated structures are wide and varied. Some use monocular images whilst others employ multiple views. Tracking in monocular images tends to be more difficult due to ambiguities arising from using just one camera view. Appearance-based approaches are often used (Holden & Owens 2003) although model based approaches have also been employed (Sminchisescu 2002).

Tracking with multiple views generally uses a model based approach since 3D information can be readily extracted. Techniques used include model fitting to volume reconstruction data obtained from multiple views (Kehl *et al.* 2005), silhouette fitting from multiple views (Carranza *et al.* 2003) and even 3D motion field reconstruction (Theobalt 2004).

*National ICT Australia Limited is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence Program

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

There are many challenges in tracking articulated structures. Fast limb movements, and self occlusion due to overlapping limbs can be problematic. Another issue is how to robustly optimise over a kinematic chain of joints. Difficulty arises when multiple joints along the kinematic chain can produce similar limb movements creating ambiguity. A commonly employed method to solve this is hierarchical tracking whereby each segment or a subset of segments from the articulated body is optimised separately, starting from the base joint (Carranza *et al.* 2003) (Kehl *et al.* 2005).

As opposed to tracking subsets of the hand one at a time which can be slow, our tracking system optimises the hand model parameters all at once. However, as shown later, doing this directly is prone to instability for harder motions such as a hand grip. This paper proposes a simple method of using optical flow information to initialise the starting step sizes for our optimisation algorithm. Applying this helps the system track gripping motions better while still allowing the hand parameters to be optimised all at once.

The paper is structured as follows; section 2 describes the tracking system with particular reference to the chosen hand model, cost function and the stochastic optimisation algorithm SMD (Schraudolph 1999). Section 3 addresses the method of incorporating optical flow information into step size initialisation. Results and discussion are presented in section 4 and 5 respectively. A conclusion is given in section 6.

2 Tracking System

A pair of calibrated SONY XCD-X710CR firewire cameras is used to acquire images of a human hand. The calibrated cameras are orientated in a convergent setup around the hand. Video sequences are captured at a resolution of 640x480 pixels and at 30 frames per second. For simplicity, a black background is used for silhouette extraction via background subtraction.

The actual tracking follows a model based approach in a stochastic optimisation framework. Points are sampled from the surface of the hand model and projected onto the image planes. Stratified sampling is used to ensure that the sample set contains points from each of the articulated segments in the hand model. The brightness constancy assumption between the pair of images and the silhouette information in each image define a cost function. Given that the projection of a 3D sample point lands on a pixel for each camera image plane, the cost for that particular sample point can be evaluated. The cost function itself will be mentioned later albeit briefly. A more detailed description of this will be reported elsewhere.

Each cost evaluation produces gradients that are

then backpropagated into the parameter space of the hand model. Given the gradients in the parameter space, the stochastic meta-descent algorithm (SMD) is used to minimise the cost. This paper will be concentrating on the step size initialisation aspect of SMD.

2.1 Hand Model

A realistic articulated hand model is used in the tracking system. The model's skeletal structure has 26 degrees of freedom (DOF) consisting of 16 joints (see figure 1). The base joint of the hand has 6 DOFs. Each digit of the hand has 4 DOFs; 2 DOFs in the MCP joints to account for the spread and bending of fingers, and 1 DOF in the PIP and DIP joint for bending. Similarly the thumb has 2 DOFs at the CMC joint, 1 DOF each for MCP and IP. Euler angles are used to parameterise the rotations for all joints.

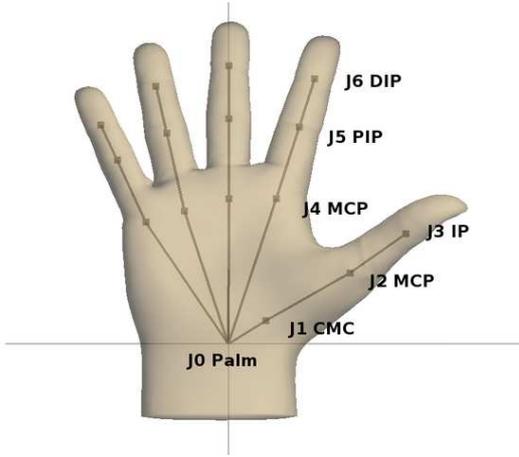


Figure 1: Hand model showing the 16 joint locations.

The skin of the hand is modeled by a detailed mesh containing approximately 9000 vertices. Linear skin blending is used to bind the skin onto the skeleton and allows for realistic deformation near joint regions (Lewis *et al.* 2000).

2.2 Cost Function

Let \mathbf{w} be the set of hand parameters ie. the Euler angles at each joint and 3 DOFs for the translation at the palm joint. The overall cost function $E(\mathbf{w})$ is evaluated by first taking N sample points from the hand model and projecting them onto the two camera image planes.

Let $p_{1,i}$ and $p_{2,i}$ be the pixel coordinates where the i th sample point of the hand model has been projected onto for camera 1 and camera 2 respectively. Assigned to each pixel coordinate $p_{j,i}$ is a brightness value $I(p_{j,i})$ and a silhouette cost value $SI(p_{j,i})$. These are used to describe the overall cost function. Note that $p_{2,i}$ and $p_{1,i}$ depend on the set of hand parameters w . Hence by the chain rule, both I and SI are dependent on \mathbf{w} .

$E(\mathbf{w})$ comprises of two parts, a silhouette cost function $E_s(\mathbf{w})$ and a cost function using the brightness constancy assumption $E_{bc}(\mathbf{w})$, each of which is computed as a sum of contributions from individual sample points, $E_s(\mathbf{w})_i$ and $E_{bc}(\mathbf{w})_i$ respectively. Let α be a scaling factor for $E_s(\mathbf{w})$. Then the overall cost function $E(\mathbf{w})$ we wish to minimise is

$$E(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{N} \sum_i^N (\alpha E_s(\mathbf{w})_i + E_{bc}(\mathbf{w})_i) \quad (1)$$

2.2.1 Silhouette

Silhouette information is used as a global constraint on the region which the projected hand model can occupy for each of the camera views.

As mentioned, the silhouette of the hand is obtained via background subtraction. A generalised distance transform is performed over the silhouette image that assigns to each pixel a value SI based on the pixel's proximity to the hand silhouette. The silhouette cost function is just SI of the pixel where the sample point on the hand is projected onto summed over $M = 2$ camera views,

$$E_s(\mathbf{w})_i = \sum_j^{M=2} SI(p_{j,i}). \quad (2)$$

2.2.2 Brightness Constancy Assumption

The brightness constancy assumption is used for local fine tuning by resolving pose ambiguities in silhouette information.

Let $I_{1,i}$ and $I_{2,i}$ be the brightness intensity at $p_{1,i}$ and $p_{2,i}$. Then the brightness constancy cost function is given as,

$$E_{bc}(\mathbf{w})_i = \frac{1}{2} \|I_{1,i} - I_{2,i}\|^2. \quad (3)$$

2.3 Optimisation Algorithm

The tracking system uses stochastic meta descent (SMD) as its optimisation algorithm (Schraudolph 1999). SMD is a stochastic gradient descent algorithm with a clever step size adaptation scheme. SMD excels as an optimisation algorithm for noisy cost functions by compromising between the robustness of first order steepest gradient descent with the fast convergence rate of second order gradient descent algorithms.

Our cost function exhibits sampling noise since only a finite set of points on the hand model is used to evaluate the cost function. Therefore the cost obtained is only an approximation of the true cost. There is also discretisation noise introduced by the evaluation of image gradients (computed using Sobel filters) and measurement noise from the camera. SMD has mechanisms to deal with these effects by taking into account the past history of step sizes in order to dampen erratic changes in the cost.

The following is a brief summary of SMD. Let \mathbf{w}_t be the vector of hand parameters at time t . In addition, let \mathbf{g}_t be the gradient of the cost function with respect to the hand parameters, and \mathbf{p}_t be the vector of step sizes. Firstly, the step size \mathbf{p}_t is updated by a meta step size scalar μ and an auxiliary vector \mathbf{v}_t in the following way:

$$\mathbf{p}_{t+1} = \mathbf{p}_t \cdot \max\left(\frac{1}{2}, \mathbf{1} + \mu \mathbf{v}_t \cdot \mathbf{g}_t\right), \quad (4)$$

where \cdot denotes the Hadamard (component-wise) product.

Then the update of the hand parameters \mathbf{w}_{t+1} is given as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{p}_{t+1} \cdot \mathbf{g}_t. \quad (5)$$

The auxiliary vector \mathbf{v}_t can be thought of as an exponential average of the past history of step sizes and also incorporates the curvature information of the cost function through the Hessian matrix \mathbf{H}_t . Let $0 \leq \lambda < 1$ be the decay factor for the exponential averaging. Then \mathbf{v}_t is updated by

$$\mathbf{v}_{t+1} = \lambda \mathbf{v}_t + \mathbf{p}_{t+1} \cdot (\mathbf{g}_t - \lambda \mathbf{H}_t \mathbf{v}_t). \quad (6)$$

Note that the hessian itself need not be calculated directly since it is only the hessian vector product $\mathbf{H}_t \mathbf{v}_t$ that is required. There are faster methods available for calculating hessian vector products (Schraudolph 2001).

Typically for the starting frame of a video sequence at iteration $t = 0$, \mathbf{w}_0 is determined by manually fitting the hand model onto the hand images. \mathbf{v}_0 is initialised to 0 and \mathbf{p}_0 is set to a uniform step size. The same procedure follows for the next video frame except \mathbf{w}_0 is carried over from the last frame. This paper focuses on the effects of setting \mathbf{p}_0 using optical flow.

2.3.1 Imposing Constraints

Constraints on the hand parameters are handled in SMD by mapping any outlying parameters back onto the feasible region (Bray *et al.* 2005):

$$\mathbf{w}_{t+1}^c = \text{constrain}(\mathbf{w}_{t+1}) \quad (7)$$

This change must be somehow reflected in \mathbf{v}_{t+1} and consequently the next step size evaluation \mathbf{p}_{t+2} . This is done by calculating the hypothetical constrained gradient \mathbf{g}_t^c given by

$$\mathbf{g}_t^c = \frac{\mathbf{w}_t - \mathbf{w}_{t+1}^c}{\mathbf{p}_{t+1}}, \quad (8)$$

and using this constrained gradient in the \mathbf{v}_t update (6). In effect, this modification completely replaces the iteration step that moves \mathbf{w}_{t+1} out of bounds with one that moves \mathbf{w}_{t+1} onto the bound.

3 Step Size Initialisation

The articulated motion of a hand can be difficult to track. Part of the difficulty lies in the relatively high dimensionality of the parameter space. A typical problem arises when multiple joints can each induce the same effect. For example, a bending movement at the tip of a finger can be the result of z-rotations of any joint along the kinematic chain. Tracking accuracy then becomes an issue of how to best balance the contributions from all possibly affecting joints.

Favoring certain joints to rotate more than the rest can be achieved by varying the step size. Intuitively, increasing the step size for a given parameter causes a larger parameter change in the parameter update, and conversely decreasing the step size inhibits parameter change. As mentioned, SMD already has an automated step size adaptation process that takes care of finding optimal step sizes over the length of iterations for a given video frame. What the results of this paper suggest is that it is helpful to initialise SMD with appropriate starting step sizes.

The rationale for using optical flow as a cue for step size initialisation is straightforward; many articulated movements of the hand such as gripping involve various segments moving at different speeds (see figure 2). A joint associated to a segment that is not moving as fast as the rest should have a reduced step size. By observing the optical flow in the scene, one can determine which parts of the hand are moving more than others and adjust the initial set of step sizes accordingly.

3.1 Optical Flow

Optical flow is used as an approximate observation of the movement at various segments of the hand and is calculated by applying the Horn & Schunk algorithm (Horn & Schunk 1981).

In this algorithm, optical flow is calculated by the minimisation of two constraint terms. One is known as the optical flow constraint equation e_c , and the other is the smoothness constraint term e_s . Let u and v be the x and y components of the optical flow in the image. In addition, let E_x, E_y, E_t be the change in brightness with respect to x, y and time t respectively. Then,

$$e_c = \int \int (E_x u + E_y v + E_t)^2 dx dy, \quad (9)$$

and

$$e_s = \int \int \left(\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right) dx dy. \quad (10)$$

Let λ_{opt} be a scaling factor. The total error e_{opt} being minimised over u and v to determine optical flow is

$$e_{opt} = e_s + \lambda_{opt} e_c \quad (11)$$

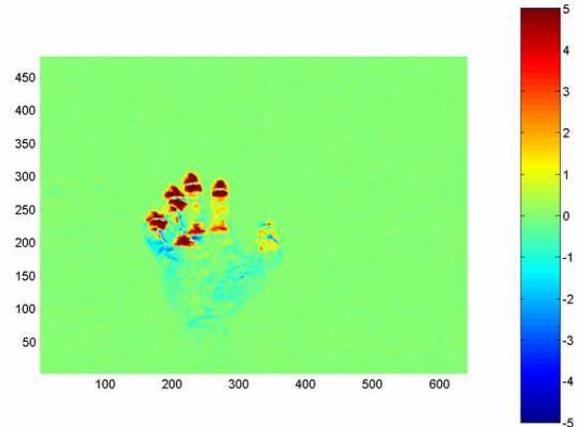


Figure 2: Differing amounts of optical flow under a gripping hand motion.

3.2 Step Size Initialisation Method

Step size initialisation requires the optical flow for each segment of the hand to be calculated. The hand model is divided into 16 (1 for the palm and 3 for each digit) segments, where each segment contains a joint. To obtain an estimate of the optical flow of a particular segment, Q sample points belonging to the segment are projected onto the camera image plane. Let $|o_i|$ be the magnitude of the optical flow vector o at the pixel where the i th sample point is projected onto. Then the optical flow O_k of a joint segment k is taken as:

$$O_k = \frac{1}{Q} \sum_i^Q |o_i|. \quad (12)$$

If the optical flow for a given segment k is below the noise threshold $T_n > 0$, then the joint is considered to be stationary and the step sizes for the parameters of that joint are decreased. Let s be the default value that each parameter step size is initially set to

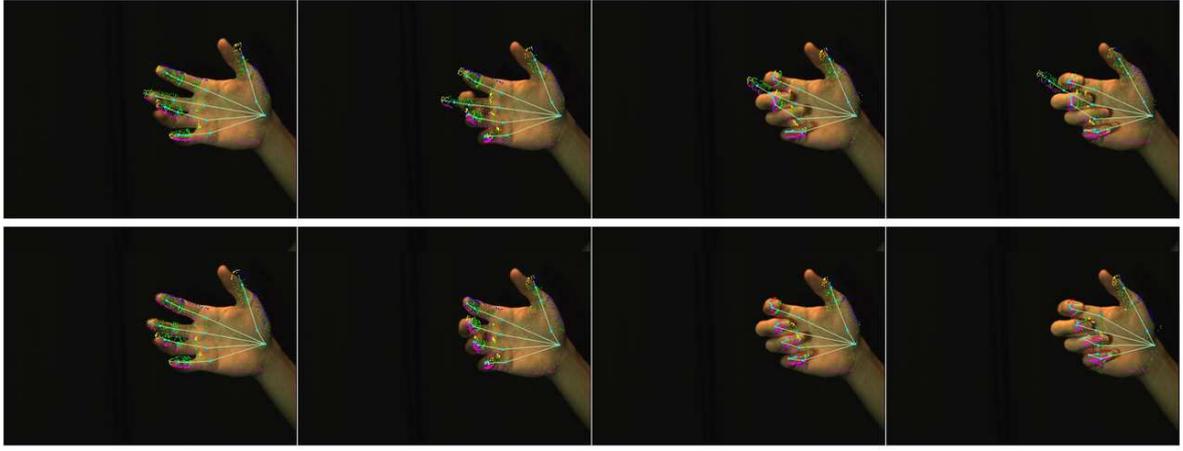


Figure 3: Gripping movement over 15 frames. (top) uniform step size (bottom) using optical flow for step size initialisation.

and let \mathbf{p}_k be the step size vector of all the parameters that control joint k . In addition, let $0 < \beta < 1$ be a scaling factor. Then,

$$\mathbf{p}_k = \begin{cases} s, & O_k > T_n \\ \beta s, & \text{otherwise} \end{cases} \quad (13)$$

The next stage in step size adjustment is comparing the optical flow in the palm with each digit. If the optical flow ratio between a digit (the optical flow of a digit is taken as the average of the optical flow of the 3 segments belonging to the digit) and the palm is higher than a certain threshold $T_f > 1$, then it is likely that the digit is going into a bending or flexing motion. Therefore one would increase the step sizes of all the parameters controlling the digit. Otherwise one would decrease the step sizes as the perceived movement of the digit is likely to be due to rotations and translations of the base joint in the palm.

Let \mathbf{p}_d be the step size vector of all the parameters that control digit d . In addition let O_d and O_p be the optical flow of digit d and the palm respectively. Then,

$$\mathbf{p}_d = \begin{cases} \mathbf{p}_d \frac{O_d}{T_f}, & \frac{O_d}{\max(O_p, T_n)} > T_f \\ \beta \mathbf{p}_d, & \text{otherwise} \end{cases} \quad (14)$$

4 Results

All experiments have been performed over a set of short video sequences under different forms of hand movements: gripping, spinning and translating movements. Each video sequence is first processed by the tracking system with a uniform step size initialisation. The initial hand pose for the first frame is estimated by eye. The α scaling factor for the silhouette cost function is set to 3.5. The SMD parameters μ and λ are set to 0.1 and 0.7 respectively. Each frame has been allowed to be optimised over a maximum of 100 SMD iterations. Approximately 150 sample points are used per iteration. Repeated testings are then used to determine an optimal step size value that is applied to all \mathbf{p}_0 entries in SMD. \mathbf{p}_0 is found to be optimal between 0.2 - 0.4.

Keeping the same settings except for \mathbf{p}_0 , the same set of experiments are repeated with step size initialisation using optical flow. The default step size s is set to 0.5 and the decrease scaling factor β is set to 0.6. T_n and T_f are set to 0.12 and 4.6 respectively.

A scaling factor $\lambda_{opt} = 0.001$ has been used for the computation of optical flow.

The effect of using optical flow for step size initialisation is most apparent for the gripping motion while there is no improvement to hand movements due to rotations or translations of the palm joint alone.

In the case of the gripping motion (see figure 3), it appears that without step size initialisation the tracker has difficulty determining the required amount of movement for each joint to follow through the motion. Errors accumulated over the frames cause the tracker to lose the pose altogether in the final frame. Whereas using optical flow for step size initialisation allows the tracker to follow the gripping motion better.

As expected for the spinning case (see figure 4), step size initialisation using optical flow does not help in resolving ambiguities due to severe self occlusion.

The effect on translating motion (see figure 5) is generally not significant, though there is worse performance at the tips of the thumb and the little finger. This is likely due to the step size for those joints being set too high.

5 Future Improvements

There is ample room for further refinements to the step size initialisation scheme. One can perhaps take into account the direction of optical flow. For example, when rotating around the x-axis for the palm joint (rotating around the x-axis causes the hand to spin in figure 4), opposing sides of the palm have opposing flow directions. Similarly, a change in depth along the camera view will induce components of opposing flow directions in each hand segment. These patterns can be best learnt in an ICA (independent component analysis) framework. In addition, step sizes along each hand digit can be refined properly by taking into account the kinematics of the linked joints of the digit. At present, we are working on creating rendered camera images of the hand by simulating hand movements in OpenGL to provide a ground truth assessment of the tracking accuracy.

6 Conclusion

The notion of utilising the motion of segments of the hand as additional information for step size initialisation has been proposed here. Optical flow has been used as an observation of this movement seen on the

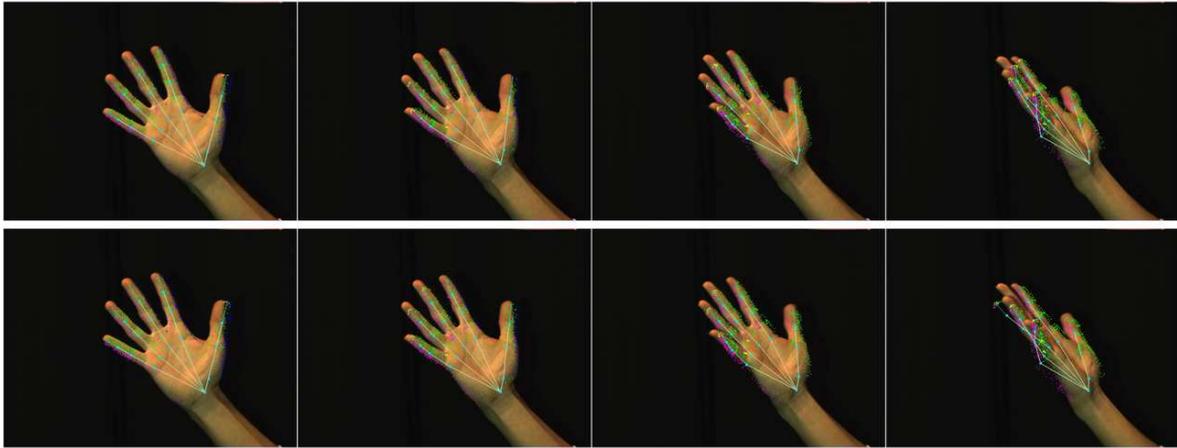


Figure 4: Spinning movement over 15 frames. (top) uniform step size (bottom) using optical flow for step size initialisation.

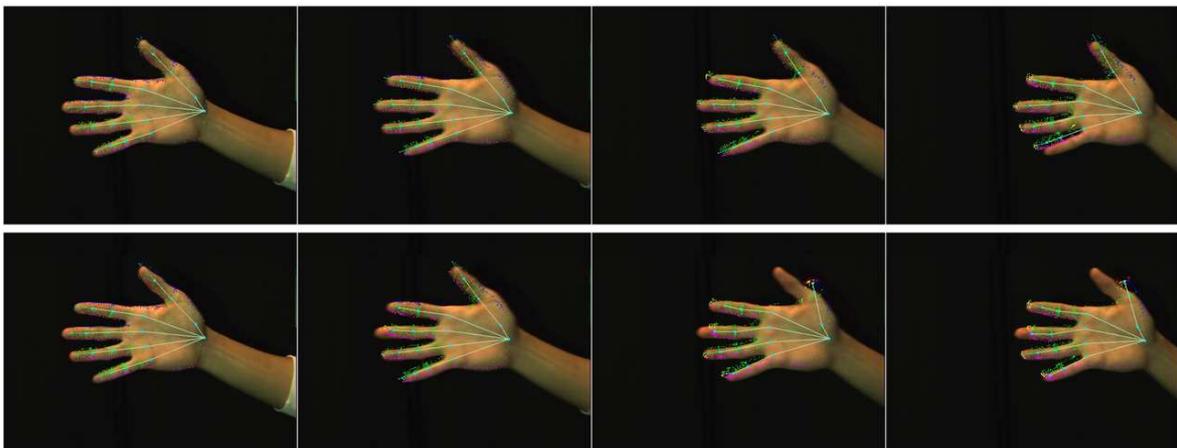


Figure 5: Translating movement over 15 frames. (top) uniform step size (bottom) using optical flow for step size initialisation.

camera image planes. Step sizes of parameters associated to a particular skeletal joint are altered by comparing the relative motion of different segments of the hand model. Despite the simplicity of the current approach to step size initialisation, using optical flow information has proved to be highly efficient in improving tracking performance for gripping motions. Future improvements are planned to refine these step size adjustments in a more vigorous manner.

7 Acknowledgments

The author would like to thank Jochen Trumpf, Nic Schraudolph, and Roland Goecke for their supervision. In addition, the author wishes to thank the BIWI computer vision laboratory, ETH, for permitting the use of their detailed hand model in this project.

References

Bray, M. , Koller-Meier, E. , Müller, P. , Schraudolph, N. & Van Gool, L. (2005), 'Stochastic Optimization for High-Dimensional Tracking in Dense Range Maps', *IEEE Proceedings on Vision, Image & Signal Processing*, Vol. 152, No. 4, pp. 501–512.

Carranza, J. , Theobalt, C. , Magnor, M. & Seidel, H.P. (2003), 'Free-Viewpoint Video of Human Actors', *ACM SIGGRAPH*, San Diego, USA, pp. 569–577.

Holden, E.J. , Lee, G. & Owens, R. (2005), 'Automatic Recognition of Colloquial Australian Sign Language', *IEEE workshop on Motion and Video Computing*, Colorado, USA, Vol. 2, pp. 183–188.

Holden, E.J. & Owens, R. (2003), 'Representing the Finger-only Topology for Hand Shape Recognition', *International Journal of Machine Graphics and Vision*, Vol. 12, No. 2, pp. 187–202.

Horn, B.K.P. & Schunck, B.G. (1981), 'Determining Optical Flow', *Artificial Intelligence*, Vol. 17, pp. 185–203.

Kehl, R. , Bray, M. & Van Gool, L. (2005), 'Full Body Tracking from Multiple Views Using Stochastic Sampling', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, USA pp.129-136

Lewis, J.P. , Corder, M. & Fong, N. (2000), 'Pose Space Deformations: A Unified Approach to Shape Interpolation and Skeleton Deformation', *ACM SIGGRAPH, Computer Graphics Proceedings*, New Orleans, USA.

- Schraudolph, N. (2001), 'Fast Curvature Matrix-Vector Products', *In Proceedings of International Conference on Artificial Neural Networks*, Vol. 2130, pp. 19–26.
- Schraudolph, N. (1999), 'Local Gain Adaptation in Stochastic Gradient Descent', *In Proceedings of International Conference on Artificial Neural Networks*, Edinburgh, Scotland, pp. 569–574.
- Sminchisescu, C. (2002), *Estimation Algorithms for Ambiguous Visual Models – Three-Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences*, PhD thesis, Institute National Polytechnique de Grenoble (INRIA), France.
- Theobalt, C. , Carranza, J. , Magnor, M. & Seidel, H.P. (2004), 'Combining 3D flow fields with silhouette-based human motion capture for immersive video', *Graphical Models*, Vol. 66, No. 6, pp. 333–351.

Hand gestures for HCI using ICA of EMG

Ganesh R Naik¹ Dinesh Kant Kumar¹ Vijay Pal Singh¹
 Marimuthu Palaniswami²

¹School of Electrical and Computer Engineering
 Royal Melbourne Institute of Technology, Melbourne, Australia

²The Department of Electrical and Electronic Engineering
 The University of Melbourne, Melbourne, Australia

Email: {ganesh.naik@rmit.edu.au, dinesh@rmit.edu.au, vijaypal@ieee.org,
 swami@ee.unimelb.edu.au }

Abstract

Aiming at the use of hand gestures for human-computer interaction, this paper presents an approach to identify hand gestures using muscle activity separated from electromyogram (EMG) using independent component analysis. While there are a number of previous reported works where EMG has been used to identify movement, the limitation of the earlier works is that the systems are suitable for gross actions, and when there is one prime-mover muscle involved. This paper reports overcoming the difficulty by using independent component analysis to separate muscle activity from different muscles and classified using backpropagation neural networks. The paper reports experimental results where the system was accurately able to identify the hand gesture using this technique for all the experiments (100%). The system has been shown not to be sensitive to electrode position as the experiments were repeated on different days. The advantage of such a system is that it is easy to train by a lay user, and can easily be implemented in real time after the initial training.

Keywords: Independent Component Analysis (ICA), Surface Electromyography (SEMG), Root Mean Square (RMS).

1 Introduction

In recent years, hand gesture recognition has become a very active research theme because of its potential use in human-Computer interaction (HCI). Identification of hand gesture has numerous human computer interface (HCI) applications related to controlling machines and computers. Some of the commonly employed modalities include vision based systems (Schlenzig, Hunter & Jain 1997, Rehg & Kanade 1994), mechanical sensors (Pavlovic, Sharma & Huang 1997), and the use of electromyogram, an indicator of muscle activity (Cheron, Draye, Bourgeois & Libert 1996, Koike & Kawato 1996). Surface Electromyogram has an advantage of being easy to record, and is non-invasive.

Surface Electromyogram (SEMG) is a result of the spatial and temporal integration of the motor unit action potential (MUAP) originating from different motor units. It can be recorded non-invasively and used for dynamic measurement of muscular function. It is typically the only in vivo functional examination

of muscle activity used in the clinical environment. The analysis of EMG can be broadly categorised into two;

- Gross and global parameters.
- Decomposition of EMG into MUAP.

Hand movement is a result of complex combination of multiple muscles. While Djuwari et al. (Djuwari, Kumar, Polus & Raghupathy 2003) have reported success in the use of multiple channels SEMG recording for the purpose, but the system is sensitive to the location of the electrodes and suitable for five discrete movements only. The cross-talk that exists due to multiple overlapping muscles in the forearm makes the system sensitive to the inter-subject variability and this problem is more significant when the muscle activation is relatively weak. To identify the movement and gesture of the hand more precisely, it is important to identify the muscle activity of each of the muscles responsible for the action. Similarity in the spectrum and other properties of the activity from the different muscles makes the separation of these difficult. There is a need to separate the muscle activity originating from different muscles. With little or no prior information of the muscle activity from the different muscles, this is a blind source separation (BSS) task.

Independent component analysis (ICA) is an iterative BSS technique that has been found to be very successful in audio and biosignal applications. ICA has been proposed for unsupervised cross talk removal from SEMG recordings of the muscles of the hand (Greco, Costantino, Morabito & Versaci 2003). Research that isolates MUAP originating from different muscles and motor units has been reported in 2004 (Nakamura, Yoshida, Kotani, Akazawa & Moritani 2004). A denoising method using ICA and high-pass filter banks has been used to suppress the interference of electrocardiogram (ECG) in EMG recorded from trunk muscles (Yong, Li, Xie, Pang, Yuzhen & Luk 2005). Muscle activity originating from different muscles can be considered to be independent, and this gives an argument to the use of ICA for separation of muscle activity originating from the different muscles. This paper proposes the use of ICA for separation of muscle activity from the different muscles in the forearm to identify the hand action.

ICA is an iterative technique where the only model of the signals is the independence, and the distribution. The outcome of ICA is that the signals are separated without there being any information of the order of the sources. While this difficulty is generally not consequential for audio signals, this would be of concern while working with muscle activity. The spatial location of the active muscle activity is the determining factor of the hand gesture. To overcome this difficulty, one approach that has been reported is

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

the use of prior knowledge of the anatomy. The advantage of this approach is the model based approach that provides a well defined muscle activity pattern. The difficulty with this approach is the need for well defined location of the electrodes.

2 Hand gesture identification for HCI

In our daily lives we interact with other people and objects to perform a variety of actions that are important to us. Computers and computerised machines have become a new element of our society. They increasingly influence many aspects of our lives. Human-computer interaction is an area concerned with the design, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.

The use of hand gesture provides an attractive alternative to cumbersome interface devices for human computer interaction applications. Human hand gestures are a mean of non verbal interactions among people. They range from simple actions of pointing at objects and moving them around to the more complex ones that express our feelings or allow us to communicate with others. The HCI interpretation of gestures requires dynamic and or/static configurations. Of the human hand, arm and sometimes, body be measurable by the machine. Hand gestures are a new mode for HCI. Visual interpretation of hand/arm movements carries a tremendous advantage over other techniques that require the use of mechanical transducers. It is not obstructive. Numerous approaches have been applied to the problem of visual interpretation of gestures for HCI. Many of those approaches have been chosen and implemented so that they focus on one particular aspect of gestures: Hand tracking, pose classification, or hand posture interpretations (Schlenzig, Hunter & Jain 1997, Rehg & Kanade 1994).

Recently a number of approaches based on hand gesture identification have been proposed for human computer interaction. Wheeler et. al. demonstrated that neuroelectric joy sticks and key boards can be used for HCI (Wheeler & Jorgensen 2003). Trejo et. al. (Trejo et. al. 2003) developed a technique for multi modal neuroelectric interface. The most recent work includes the investigation of eleven normally limbed subjects (eight males and four females) for six distinct limb motions: wrist flexion, wrist extension, supination, pronation, hand open, and hand close. Each subject underwent four 60-s sessions, producing continuous contractions (Chan & Englehart 2005).

A number of efficient solutions to gesture input in HCI exist, such as:

- Restrict the recognition situation.
- Use of input devices (e.g. data glove).
- Restrict the object information.
- Restrict the set of gestures.

In traditional HCI, most attempts have used some device, such as an instrumented glove, for incorporating gestures into the interface. If the goal is natural interaction in everyday situations this might not be acceptable. Vision based approach to hand-centered HCI has been proposed in recent years. However, a number of applications of hand gesture recognition for HCI exist, using Computer vision technique. Mostly they require restricted backgrounds and camera positions, and a small set of gestures, performed with one hand (Pavlovic, Sharma & Huang 1997).

In this report we propose Hand gesture identification which uses the prior knowledge of muscle

anatomy. This is a model based approach that provides a well defined muscle activity pattern.

3 Surface Electromyogram

SEMG is a non-invasive recording of the muscle activity and finds application in sports training, rehabilitation, machine and computer control, occupational health and safety, and for identifying posture disorders. There is a near linear relationship between RMS of SEMG and the finger flexion-extension - suggesting the use of SEMG for bio-control for anthropomorphic tele-operators and Virtual Reality entertainment (Gupta & Reddy 1996). There is useful information of the posture from the muscle activity of the lumbar muscles. SEMG amplitude and frequency have been investigated as indicators of localized muscular fatigue. Amplitude and spectral information of EMG have also been exploited to estimate force of muscle contraction and torque (Moritani & Muro 1987). These applications require automated analysis and classification of SEMG.

SEMG may be affected by various factors such as:

- The muscle anatomy (number of active motor units, size of the motor units, the spatial distribution of motor units).
- Muscle physiology (trained or untrained, disorder, fatigue).
- Nerve factors (disorder, neuromuscular junction).
- Contraction (level of contraction, speed of contraction, isometric/non-isometric, force generated).
- Artefacts (crosstalk between muscle, ECG interference).
- Recording apparatus factors (recording-method, noise, electrode's properties, recording sites).

The anatomical/ physiological processes such as properties and dimensions of tissues, and force and duration of contraction of the muscle are known to influence the signal. SEMG is also influenced by onset of muscle fatigue, and contraction of other muscles in the close vicinity. Each of the factors can be used as a criterion to categorise the input signal. One property of the SEMG is that the signal originating from one muscle can generally be considered to be independent of other bioelectric signals such as electrocardiogram (ECG), electro-oculargram (EOG), and signals from neighbouring muscles. This opens an opportunity of the use of independent component analysis (ICA) for this application.

4 Basic Principles of Independent Component Analysis (ICA)

It is often required to separate the original signals from the mixture of signals, when there is little information available of the original signals and there is an overlap of the signals in time and frequency domain. Even if there is no or limited information available of the original signals or the mixing matrix, it is possible to separate the original signals using independent component analysis (ICA) under certain conditions. ICA is an iterative technique that estimates the statistically independent source signals from a given set of their linear combinations. The process involves determining the mixing matrix. The independent sources could be audio signals such as speech, voice, music, or signals such as bioelectric signals.

Independent Component Analysis is a technique for extracting statistically independent variables from a mixture of them. ICA searches for a linear transformation to express a set of random variables as linear combinations of statistically independent source variables (Comon 2001). The criterion involves the minimization of the mutual information expressed as a function of high order cumulants. ICA separates signals from different sources into distinct components. The technique is based on unsupervised learning rules where reduction of mutual information and increase in Gaussianity are the cost functions. Given a set of multidimensional observations, which are a result of linear mixing of unknown independent sources through an unknown mixing source, ICA can be employed to separate the signals from the different sources. The independent sources may be sources for audio signals such as speech, voice, music, or signals such as bioelectric signals. If the mixing process is assumed to be linear, it can be expressed as

$$x = As \quad (1)$$

where $x = (x_1, x_2, \dots, x_n)$ is the recordings, $s = (s_1, s_2, \dots, s_n)$ the original signals and A is the $n \times n$ mixing matrix of real numbers. This mixing matrix and each of the original signals are unknown. To separate the recordings to the original signals, an ICA algorithm performs a search of the un-mixing matrix W by which observations can be linearly translated to form Independent output components so that

$$s = Wx \quad (2)$$

For this purpose, ICA relies strongly on the statistical independence of the sources s . This technique iteratively estimates the un-mixing matrix using the maximisation of independence of the sources as the cost function (Hyvarinen, Karhunen & Oja 2001).

The success of ICA to estimate independent sources is dependent on the fulfillment of the following conditions.

- The sources must be statistically independent.
- The sources must have non Gaussian distributions. However, ICA can still estimate the sources with small degree of non-Gaussianity
- The number of available mixtures N must be at least the same as the number of the independent components M .
- The mixtures must be (can be assumed as) linear combination of the independent sources.
- There should be no (little) noise and delay in the recordings.

ICA also suffers from the following unavoidable ambiguities.

- The order of the independent components cannot be determined (it may change each time the estimation starts)
- The exact amplitude and sign of the independent components cannot be determined.

There are several estimation algorithms for the ICA technique. FastICA algorithm is based on negentropy (negative entropy) and has been developed and proposed by the team at the Helsinki University of Technology (FastICA 2005). This algorithm uses negentropy as a measure of non-Gaussianity of the signals and uses fixed point iteration scheme. It is faster than conventional gradient descent scheme. This paper reports the use of FastICA for analysis.

4.1 ICA for SEMG applications

A number of researchers have reported the use of ICA for separating the desired SEMG from the artefacts and from SEMG from other muscles. While details differ, the basic technique is that different channels of SEMG recordings are the input of ICA algorithm. ICA has also been used by Heido et al. (Nakamura, Yoshida, Kotani, Akazawa & Moritani 2004) to decompose the SEMG recordings in terms of the MUAPs. In their paper, they have acknowledged the drawbacks and the necessary conditions required for the success of the ICA, but have not demonstrated how the suitability of their experimental data for ICA application. With the help of 8 channel recordings, the SEMG signal has been decomposed into MUAPs that may have originated from large number of motor units. This could make the number of sources to be more than the number of recordings, making it unsuitable for standard ICA.

The fundamental principle of ICA is to determine the un-mixing matrix and use that to separate the mixture into the independent components. The independent components are computed from the linear combination of the recorded data. The success of ICA to separate the independent components from the mixture depends on the properties of the recordings.

4.2 Statistical Properties of SEMG Recordings

ICA uses the Gaussianity of the signals as a cost function to generate the un-mixing matrix and hence signals that have Gaussian distribution are unsuitable for ICA applications (Hyvarinen, Karhunen & Oja 2001). Mathematical manipulation demonstrates that all matrices will transform this kind of mixtures to another Gaussian data. However, a small deviation of density function from Gaussian may make it suitable as it will provide some possible maximization points on the ICA optimization landscape, making Gaussianity based cost function suitable for iteration. If one of the sources has density far from Gaussian, ICA will easily detect this source because it will have a higher measure of non Gaussianity and the maxima point on the optimization landscape will be higher. If more than one of the independent sources has non-Gaussian distribution, those with higher magnitude will have the highest maxima point in the optimization landscape. Given a few signals with distinctive density and significant magnitude difference, the densities of their linear combinations will tend to follow the ones with higher amplitude. Since ICA uses density estimation of a signal, the Components with dominant density will be found easier.

Signals such as SEMG have probability densities that are close to Gaussian while artefacts such as ECG and motion artefacts have non Gaussian distributions. From the above, it can be suggested that ICA may suitably isolate some of the above signals, while its efficacy for separating the others maybe questionable. It is difficult to identify the quality of separation of EMG from one muscle and the neighbouring muscles, making it difficult to confirm or negate the above. This paper reports the use of ICA to separate EMG from different muscles. As the signal properties of EMG are close to Gaussian, and there is no information available of the original signal, the only measure of quality possible is to determine the accuracy of the system to identify the hand gesture correctly.

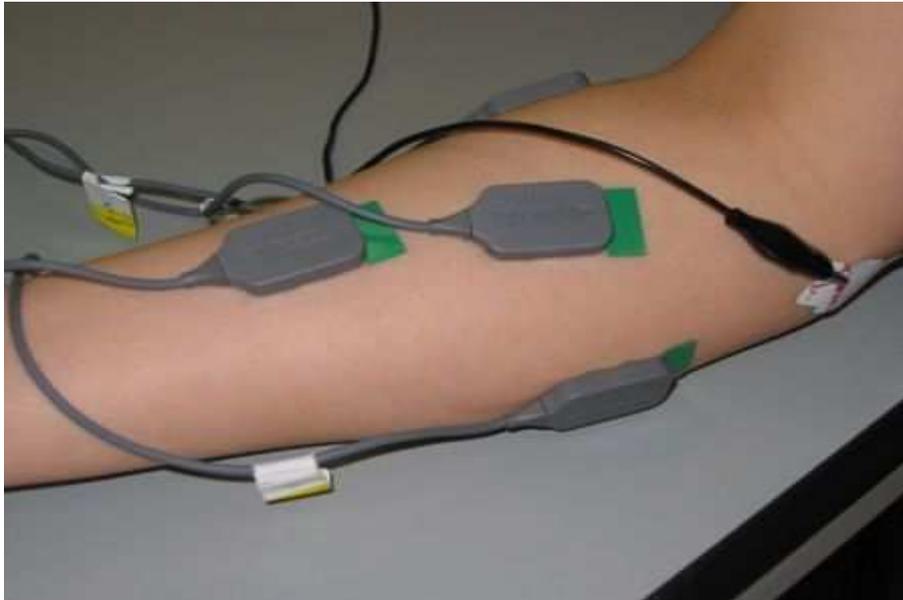


Figure 1: Placement of Electrodes

5 Methodology

5.1 Experimental Procedure

RMIT University ethics committee granted approval to conduct experiments on human subjects and acquire Surface EMG using surface electrodes. For the data acquisition a proprietary SEMG acquisition system by Delsys (Boston, MA, USA) was used. Four electrode channels were placed over four different muscles as indicated in the table 1 and figure 1. Each channel is a set of two differential electrodes with a fixed inter-electrode distance of 10mm and a gain of 1000 which is shown in figure 2. Before placing the electrodes subject's skin was prepared by lightly abrading with skin exfoliator to remove dead skin that helps in reducing the skin impedance to less than 60 Kilo Ohm. Skin was also cleaned with 70% v/v alcohol swab to remove any oil or dust on the skin surface.

Channel	Muscle	Function
1	Brachioradialis	Flexion of forearm
2	Flexor Carpi Radialis (FCR)	Abduction and flexion of wrist
3	Flexor Carpi Ulnaris (FCU)	Adduction and flexion of wrist
4	Flexor Digitorum Superficialis (FDS)	Finger flexion while avoiding wrist flexion

Table 1: Placement of the Electrodes over the skin of the forearm

ICA is suitable when the numbers of recordings are same as or greater than the number of sources. This paper reports using 4 channels of EMG recorded during hand actions that required not greater than 4 independent muscles. This ensures that the unmixing matrix is a square matrix of size of 4 x 4.

The experiments were repeated on two different days. Subject was asked to keep the forearm resting on the table with elbow at an angle of 90 degree in a comfortable position. Three hand actions were performed and repeated 12 times at each instance. Each time raw signal sampled at 1024 samples/second was recorded. A suitable resting time was given be-

tween each experiment. There was no external load. The actions were complex to determine the ability of the system when similar muscles are active simultaneously and are listed below:

- Wrist flexion (without flexing the fingers).
- Finger flexion (ring finger and the middle finger together without any wrist flexion).
- Finger and wrist flexion together but normal along centre line.

While Brachioradialis is an elbow flexor, a very little activity may be recorded in this muscle while finger and/or wrist flexion. FCU and FCR are the two wrist flexors. FDS performs the flexion of the middle finger and the ring finger.

The hand actions and gestures represented low level of muscle activity. The hand actions were selected based on small variations between the muscle activities of the different digitus muscles situated in the forearm. The recordings were separated using ICA to separate activity originating from different muscles and used to classify against the hand actions.

5.2 Analysis

The aim of this experiment was to test the use of ICA for separation of the EMG signals for the purpose of identifying hand gestures and actions.

For each hand movement we recorded 12 repetitions, lasting approximately 2.5 seconds each. The sampling rate was 1024 samples per second, this gives approximately 2500 samples. For the first set of experiments recorded signals x were analysed using matlab software package. There were four channel (recordings) electrodes and four active muscles associated with the hand gesture, this formed 4X4 mixing matrix. For each set of experiments the EMG data was analyzed using fast ICA matlab package which has been developed and proposed by the team at the Helsinki University of Technology (FastICA 2005). The mixing matrix A was computed for the first set of data only. This was kept constant throughout the experiment.

$$x = As \quad (3)$$

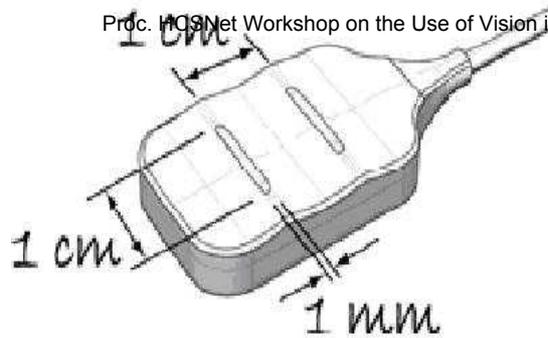


Figure 2: Electrodes. (source: www.delsys.com)

where x is the recorded data, A is the mixing matrix and s is the sources. The independent sources of motor unit action potentials that mix to make the EMG recordings were computed using the following.

$$s = Bx \quad (4)$$

where B is the inverse of the mixing matrix A . This process was repeated for each of the three hand gesture experiments. Four sources were obtained for each experiment. After separating the four sources s_a, s_b, s_c and s_d each are also 2500 samples long. Root Mean Squares (RMS) was computed for each separated sources using the following formula

$$s_{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^N s_i^2} \quad (5)$$

Where s is the source and N is the number of samples ($N = 2500$). This resulted in one number representing the muscle activity for each channel for each hand action. Hence we obtained four RMS values every time

The examples of one set of RMS values obtained during wrist flexion experiment are shown in the table 2 below.

Source	RMS(Root Mean Square)values
Source1(s1)	1.2214
Source2(s2)	1.1205
Source3(s3)	1.1846
Source4(s4)	1.2104

Table 2: Example of one set of experiment results showing the RMS (Root Mean Square) values during the wrist flexion action

RMS of muscle activity of each source represents the muscle activity of that muscle and is indicative of the force of contraction generated by each muscle. A combination of the activity from each of these muscles is responsible for the muscle activity (gesture) and has been used to identify the hand gesture. While ICA has the order ambiguity shortcoming, but by using a constant un-mixing matrix (B) for each of the experiments, the data classification can be achieved against the movement.

The above process was repeated for all three different hand actions. The outcome of this was 12 set of examples, each example pertaining to three actions. These 12 sets of examples were used to train a back-propagation neural network with 4 inputs and 3 outputs (The 4 RMS (Root Mean Square) values of the muscles were the input and the 3 RMS (Root Mean Square) values were the output). In the first part of

the experiment, RMS values of 3 recordings for subject were used to train the ANN classifier with back propagation learning algorithm. In the second part of the experiment, the neural network was trained using the data from the subject and tested similarly. The architecture of the ANN consisted of two hidden layers and the 20 nodes for the two hidden layers were optimized iteratively during the training of the ANN. Sigmoid function was the threshold function and the type of training algorithm for the ANN was gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima. In the testing section, the trained ANNs were used to classify the RMS values of recordings that were not used in the training of the ANN to test the performance of the proposed approach. The ability of the network to correctly classify the inputs against known hand actions were used to determine the efficacy of the technique.

6 Results and observations

Backpropagation neural network with 3 inputs and 4 outputs are conducted for three types of hand gestures. The result of the use of these normalized values to train the ANN using data from individual subjects showed easy convergence. The results of testing the ANN to correctly classify the test data based on the weight matrix generated using the training data is tabulated in table 3. five set of experiments. The accuracy was computed based on the percentage of correct classified data points to the total number of data points. The classification accuracy was 100% for all the experiments.

Action Performed	Action identified for experiments				
Wrist flexion	100%	100%	100%	100%	100%
Finger flexion	100%	100%	100%	100%	100%
Finger flexion & Wrist flexion	100%	100%	100%	100%	100%

Table 3: Neural network testing results

7 Discussions and Conclusion

A new approach that combines semi-blind ICA along with neural networks was used to separate and identify hand gestures. The results demonstrate that the technique can be effectively used to identify hand gestures based on surface EMG when the level of activity is very small. The authors would like to mention that

this is early stage of the work, and work needs to be done to identify inter-day variations. It is also important to test the technique for different actions, and for a large group of people. Further, there is need to automate the semi-blind operation.

8 Acknowledgement

The authors would like to thank Waichee Yau for the help with neural net work part and Vijay Pal for the conduction of the experiments. Authors are also would like to extend their gratitude to the anonymous reviewers for their helpful comments.

References

- Chan, A. D. C. & Englehart, K. B. (2005), 'Continuous myoelectric control for powered prostheses using hidden Markov models', *IEEE Transactions on Biomedical Engineering*, Vol. 52, No. 1, pp. 121-124.
- Cheron, G., Draye, J., Bourgeois, M. & Libert, G. (1996), 'A Dynamic Neural Network Identification of Electromyography and Arm Trajectory Relationship During Complex Movements', *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 5, pp. 552-558.
- Comon, P. (2001), *Independent Component Analysis, a new concept*, John Wiley, New York.
- Djuwari, D., Kumar, D. K., Polus, B. & Raghupathy, S. (2003), 'Multi-step independent component analysis for removing cardiac artefacts from back semg signals', *8th Australian and New Zealand Intelligent Information Systems Conference*, Australia.
- FastICA (2005), The FastICA MATLAB package, Helsinki University of Technology, <http://www.cis.hut.fi/projects/ica/fastica/>. Accessed 20 Feb 2006.
- Greco, A., Costantino, D., Morabito, F. C. & Versaci, M. A. (2003), 'A Morlet wavelet classification technique for ICA filtered SEMG experimental data', *Neural Networks Proceedings of the International Joint Conference*, Vol. 1, No. 2, pp. 166-171.
- Gupta, V. & Reddy, N. P. (1996), 'Surface electromyogram for the control of anthropomorphic teleoperator fingers', *Student Health Technology Information*, Vol. 29, pp. 482-487.
- Hyvarinen, A., Karhunen, J. & Oja, E. (2001), *Independent Component Analysis*, John Wiley, New York.
- Koike, Y. & Kawato, M. (1996), 'Human Interface Using Surface Electromyography Signals', *Electronics and Communications*, Vol. 79, No. 9, pp. 15-22.
- Moritani, T. & Muro, M. (1987), 'Motor unit activity and EMG power spectrum during increasing force contraction', *Eur. J. Appl. Physio. Occup.*, Vol. 56, pp. 260-265.
- Nakamura, H., Yoshida, M., Kotani, M., Akazawa, K. & Moritani, T. (2004), 'The application of independent component analysis to the multi-channel surface electromyographic signals for separation of motor unit action potential trains', *Journal of Electromyography and Kinesiology*, Vol. 14, No. 4, pp. 423-432.
- Pavlovic, V. I., Sharma, R. & Huang, T. S. (1997), 'Visual interpretation of hand gestures for human-computer interaction', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 677-695.
- Rehg, J. M. & Kanade, T. (1994), 'Vision-based hand tracking for human-computer interaction', *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 16-22.
- Schlenzig, J., Hunter, E. & Jain, R. (1994), 'Vision based hand gesture interpretation using recursive estimation', *Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, Vol. 2, pp. 1267-1271.
- Trejo, L. J., Wheeler, K. R., Jorgensen, C. C., Rosipal, R., Clanton, T. S., Matthews, B., Hibbs, A. D., Matthews, R. & Krupka, M. (2003), 'Multimodal neuroelectric interface development', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 11, No. 2, pp. 199-204.
- Wheeler, K. R. & Jorgensen, C. C. (2003), 'Gestures as input: Neuroelectric joysticks and keyboards', *IEEE Pervasive Computing*, Vol. 2, No. 2, pp. 56-61.
- Yong, Hu., Li, X., H. Xie, X., B. Pang, L., Y. Yuzhen, Cao. & Luk, K. D. K. (2005), 'Applying Independent Component Analysis on ECG Cancellation Technique for the Surface Recording of Trunk Electromyography', *IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China.

Nuisance Free Recognition of Hand Postures Over a Tabletop Display

João Carreira

Paulo Peixoto

Institute of Systems and Robotics
University of Coimbra
Polo II - Pinhal de Marrocos
3030 Coimbra
Email: {joaoluis, peixoto}@isr.uc.pt

Abstract

This paper proposes a new approach to shape classification that is well suited to the specific challenges of vision-based hand posture recognition in a multi-user tabletop collaboration scenario. We use a representation of the 2-D hand silhouette where in-plane rotation and mirror symmetry appear as particular cases of permutations, and then show how to take advantage of this pattern to develop an efficient version of the permutation invariant SVM. Invariance to these transformations is very important because the users stand around the table, and a video camera captures the scene from the top. We also report experimental results that compare this approach favorably over common classification approaches, under the stated requirements.

Keywords: tabletop interaction, vision-based gesture recognition, support vector machines

1 Introduction

Tabletop displays have been a subject of considerable interest by the Human Computer Interaction community over the last fifteen years, as they present a natural medium for computer-assisted local collaboration between people. Computer Vision could be an important sensing technology for these systems, once it gets more stable: many users already have webcams which are cheap, easily deployable, and could be used to capture hand gestures at high frequencies. Also, LCD and Plasma displays are becoming larger and more economic, and cameras can adapt seamlessly to capture different screen areas. Tabletop systems present, however, very characteristic requirements to a gesture recognition software: full rotation invariance, because the users are around the table, and mirror symmetry invariance, to equally recognize left and right hand gestures. It should also be computationally cheap enough to cope with capturing multiple users' gestures simultaneously in real time and still allow the computer to run its applications. To understand these requirements consider the setups of the applications "Room Planner" (Wu

& Balakrishnan 2003), and "CollabDraw" (Morris & Winograd 2006), illustrated in figure 1. These and other recent multi-user applications rely on newly developed multi touch sensitive tables, as Diamond Touch (Dietz & Leigh 2003), and use video projectors to provide the image. Although these sensors make a robust and dependable interface they're not without limitations, namely they're very expensive and can only directly capture information about the shape of the pressing areas of the hand against the table.

The focus of this paper is a mid-level vision problem: shape classification. We present a simple adaptation of a recent machine learning algorithm, the permutation invariant Support Vector Machine, or pi-SVM, (Shivaswamy & Jebara 2006), that combined with a properly coded representation of the silhouette of the hand, turns out to be an approach well suited to the specific problems of hand posture recognition in vision-based tabletop interfaces. In particular, it only distinguishes hand postures which differ intrinsically in shape, ignoring what we consider nuisance parameters: in-plane rotation and mirror symmetry. The approach doesn't require manual annotation of landmarks, although it requires labeled training data. The general idea of the pi-SVM is to, during training time, to transform the data in a way that both minimizes the radius of the hypersphere enclosing the points, and maximizes the margin between the points of the different classes, so that the nuisances get optimized away. Then, at test time, the transformation that best discriminates a pattern is first applied, followed by classification using a SVM. We show that in the desired tabletop setup, the nuisance parameters, to which the classification should be invariant, come up as restricted kinds of permutations, that can be handled by a less ambitious version of the pi-SVM. One that is also much more efficient than the original.

We assume there's some segmentation process that provides us with a closed contour of the hands. In our case we have been using skin color detection, which albeit improper for general situations, in the particular scenario where we're controlling the image in the LCD display it's a feasible solution.

The structure of this paper is as follows: we discuss related work in section 2, then the feature used is introduced in section 3, and customizations of the permutation invariant SVM are proposed in Section 4. Experimental results are shown, and discussed in section 5 and finally conclusions are presented in section 6.

2 Related Work

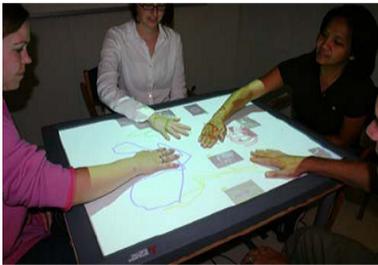
Most vision-based interfaces over a table have relied solely on fingertip tracking. For example, Letessier & Bérard (2004) matched a circular template to binarized images in order to detect fingertips independently of the orientation of the hand. Baraldi, Bimbo,

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Research described in the paper was financially supported by FCT under grant No. POSC/EEA-SRI/61451/2004. The first author was financially supported by FCT PhD grant SFRH/BD/24295/2005.



(a) Room Planner



(b) CollabDraw

Figure 1: Examples of multi-user tabletop interaction setups.

Landucci & Valli (2006) used the same method, and built a simple rule-based classifier that can discriminate between three postures, based on the number of stretched fingers. Sato, Kobayashi & Koike (2000) use an infrared camera tuned to the human temperature to segment the forearm, and then uses the principal axis of the resulting blob to guide a normalized correlation search for the fingertips. This simplicity is a result of pragmatic thinking, as ambitious approaches from more traditional vision-based gesture recognition research don't work reliably and fast enough. A good review of these can be found in work by Derpanis (2004), which divides existing approaches as model-based, appearance-based and feature-based.

The types of invariances we seek have been mostly tackled using feature-based approaches, by pursuing representations of the features that directly incorporate them. A popular feature is the boundary of silhouettes, which has no internal holes or markings, making it easily representable in 1-D, parameterized by arc length. This kind of feature is incorporated in the MPEG-7 standard, and there's a large pool of solutions developed. There are representations of this feature that present some kinds of invariances, like Fourier Descriptors with respect to rotation. A more flexible feature is the Shape Context (Belongie, Malik & Puzicha 2002), which can represent a shape with inner markings, making it possible to use directly the output of edge extractors. The authors describe a way to achieve rotation invariance using this feature, but point out that it relies on contour tangents, which are highly sensitive to noise. Other flexible approach is the use of local invariant features that represent shape key points, as the SIFT feature (Lowe 2004), which can be directly calculated from the output of a low level interest point detector. The problem of this approach, as pointed out by Belongie *et al.* is that it sacrifices the shape information available in smooth portions of object contour, and that some objects - e.g. circles - don't even have any key points.

There's also work done on incorporating invariances directly into classification algorithms. Scholkopf & Smola (2002) identify three different approaches in the context of kernel methods: generating virtual support vectors, constructing invariance kernels and jittering support vectors. The first con-

sists in generating virtual examples from the support vectors - informally speaking, the examples that are most difficult to classify - and then retraining using the new data. The virtual examples result from the application of transformations which we know *a priori* that shouldn't change the label of the example - the invariances. Invariance kernels work by directly regularizing the hyperplane in a way that trades off margin for parallelism to the directions of invariance. Finally, jittering support vectors works by transforming the example vectors in a way that the euclidean distances between them in feature space are minimal.

The problem with invariance kernels is that they can only be applied to smooth transformations. The virtual examples approach can become slow during classification time because of the increase in support vectors, and jittering can generate kernels that aren't positive definite, and so, the algorithms may not converge. A recently proposed method consists in cleaning up and reconstructing the data before training the classifier. This is the subject of the works by Bi & Zhang (2004) and Shivaswamy & Jebara (2006), with the latter resulting in the permutation invariant SVM. In order to make this paper more self-contained we introduce this algorithm in the next section, following the original presentation from Shivaswamy & Jebara (2006), while adding some additional comments from our analysis.

2.1 Permutation Invariant SVMs

The permutation invariant SVM is a binary classification algorithm, motivated by an important result in statistical learning theory (Vapnik 1995), which states that the expectation of the classification error probability is bounded by the ratio of the squared radius of the minimum hypersphere that encloses the data, to the square of the margin that separates the data points from both classes. This suggests a strategy to reduce the influence of unknown nuisance parameters in the data, by transforming the data along the desired invariances, and selecting the transformations that make the different classes most well separated, while being enclosed in a hypersphere with small radius.

In Shivaswamy & Jebara (2006) the targeted invariance was general permutation of the feature vector elements, and to this end the authors proposed calculating the radius and center of the minimal hypersphere enclosing the data, the maximal margin of the optimal separating hyperplane, and then, for each input sample, setting up a matrix of costs that indicates how favorable each different permutation is. The best permutation for each sample is then chosen by solving a Linear Assignment Problem (Papadimitriou 1982), or LAP, which can be done efficiently using the Kuhn-Munkres algorithm, also known as the Hungarian algorithm. After transforming all samples, the radius and center of the minimal hypersphere and the margin of the new optimal hyperplane are calculated again, and the rest of the process is repeated. After a number of iterations, the classifier corresponding to the optimal hyperplane of the transformed data is stored, in order to be used during test time. The training algorithm is described in Algorithm 1.

Computing the hyperplanes and hyperspheres can be done by solving the following optimization problems, present in most textbooks about kernel methods (Scholkopf & Smola 2002, Shawe-Taylor & Cristianini 2004). Let \mathbf{w} be the vector of parameters of the hyperplane that separates two classes of data samples \mathbf{x}_i , with labels y_i , and ξ_i be slack variables for accounting noise, or non-separability of the classes.

Then the maximal margin hyperplane can be estimated from the solution of the following quadratically constrained quadratic program formulation:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (1)$$

subject to

$$y_i \mathbf{w}' \mathbf{x}_i + b \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall 1 \leq i \leq n \quad (2)$$

Similarly, the centroid and radius of the smallest hypersphere enclosing the data points can be estimated from:

$$\min_{\mathbf{c}, R, \xi} R^2 + C \sum_{i=1}^n \xi_i \quad (3)$$

subject to

$$\|\mathbf{c} - \mathbf{x}_i\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0 \quad \forall 1 \leq i \leq n \quad (4)$$

In both cases the parameter C controls how acceptable it is for the margin and hypersphere radius to be violated, in order to account for noise. Large C corresponds to a hard margin and to the hypersphere containing all points.

For clarity's sake we believe it's useful to discuss the less obvious step of Algorithm 1: step 3. The idea is to find the permutation matrix that best transforms the feature vector, both in terms of how close it gets to the center of the hypersphere and to how far away it gets from the separating hyperplane. Consider the first term of the sum: if \mathbf{w} is fixed, the dot product $\mathbf{w}' \mathbf{x}$ is proportional to the distance of \mathbf{x} to the hyperplane, and that's what we want to maximize, by an appropriate choice of the permutation matrix A . If that was the only thing to optimize, the algorithm would proceed to find A by solving the maximization version of the LAP. Let $\mathbf{x}' = [x_1 x_2]$ and $\mathbf{w}' = [w_1 w_2]$. Then the reward matrix is:

$$\mathbf{w} \mathbf{x}' = \begin{bmatrix} w_1 x_1 & w_1 x_2 \\ w_2 x_1 & w_2 x_2 \end{bmatrix}$$

The LAP, with this reward matrix, amounts to finding the one to one assignment of elements of \mathbf{w} to elements of \mathbf{x} such that their sum is maximal. This \mathbf{x} effectively maximizes $\mathbf{w}' \mathbf{x}$.

Conversely the aim of the second term is to minimize the dot product $\mathbf{c}' \mathbf{x}$. This corresponds to finding the permuted \mathbf{x} which is closest to \mathbf{c} . For this to be true, \mathbf{x} should be normalized to fixed length, and have positive elements (so that the minimal angle with \mathbf{c} corresponds to a minimal distance, because $\mathbf{c}' \mathbf{x} = \|\mathbf{c}\| \|\mathbf{x}\| \cos \alpha$).

Finally, the λ parameter determines the trade off between optimizing the margin and the radius of the data enclosing hypersphere.

During test time, in order to predict the label of a test datum, the algorithm solves again the LAP problem for two different reward functions, $\lambda \mathbf{w} \mathbf{x}' - \mathbf{c} \mathbf{x}'$ and $-\lambda \mathbf{w} \mathbf{x}' - \mathbf{c} \mathbf{x}'$, getting in general two different permutations as solutions. The label corresponding to the largest absolute reward is then selected.

3 The Feature

The silhouette (S) of a segmented hand region, like the one depicted in 2, is a finite set of N_i points on the image, that define the basic shape of the hand (figure 3):

Algorithm 1 Algorithmic description of the original Permutation Invariant SVM

Input: Training data set - $(x_i, y_i)_{i=1}^n$, Maximum Iterations - max , Parameter - λ

Output: Hyperplane - (w, b) and Centroid - c

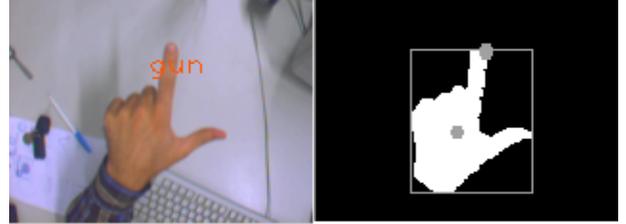
0. Set $j \leftarrow 1$

1. Solve (3) from $(x_i, y_i)_{i=1}^n$ to find centroid c^j and the radius R .

2. Solve (1) from $(x_i, y_i)_{i=1}^n$ to find (w^j, b^j) and margin M .

3. Solve Kuhn-Munkres Algorithm with reward matrix $\lambda y_i w^j x'_i - c^j x'_i$ for each i , let the permutation matrix obtained be A^{ij} .

4. If $j = max$ return (w^j, b^j, c^j) else $j \leftarrow j + 1$



(a) A posture.

(b) Result of the segmentation.

Figure 2: Contour extraction using color segmentation.

$$S = \{s_k = (x_k, y_k), k = 1, \dots, N_i\} \quad (5)$$

We assume that the silhouette S has the following properties:

- S is closed, i.e. s_1 is next to s_{N_i} .
- S has a depth of one single point (it's one dimensional).
- S is defined by accounting points in the clockwise direction.

The starting point of the definition of the representation, that we shall call a signature, is the calculation of the polar coordinates of each point s_k belonging to the contour of the segmented blob. The polar coordinates are defined in such a way that the origin of the coordinate system is the centroid $C = (c_x, c_y)^T$ of the segmented region R , defined as:

$$c_x = \frac{\sum_x \sum_y f(x,y)x}{\sum_x \sum_y f(x,y)}, \quad \text{and} \quad c_y = \frac{\sum_x \sum_y f(x,y)y}{\sum_x \sum_y f(x,y)}$$

$$\text{with } f(x,y) \begin{cases} 1 & \text{if } x,y \in R \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Given the silhouette $S = (s_1, s_2, \dots, s_{N_i})^T$ from the segmented hand on frame i we can compute the coordinates ρ_k , that corresponds to the Euclidean distance of each point to the centroid of the segmented hand blob, and θ_k , the angle:

$$\rho_k = \|s_k - C\| = \sqrt{(x_k - c_x)^2 + (y_k - c_y)^2} \quad (7)$$

$$\theta_k = \arctan \frac{(y_k - c_y)}{(x_k - c_x)}, \quad \text{with } k = 1..N_i \quad (8)$$

Having the polar coordinates, we split the silhouette S into r radial segments of equal size, and for

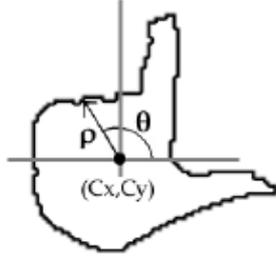


Figure 3: The silhouette's signature is defined by dividing the contour in fixed radial segments. The largest magnitude point in every segment is chosen. The center of the coordinate system is the centroid of the hand's blob.

each one we select the largest magnitude ρ'_k whose corresponding θ_k belongs to the angle interval that defines the segment. This way the signature has a fixed length of r elements. This signature is intrinsically invariant to translation of the hand in the image frame, since the silhouette is defined in relation to a coordinate system with its origin at the centroid of the hand's blob. The same is not true for scale: different distances from the camera to the hand will imply different silhouette amplitudes. A simple solution is very effective nonetheless: we normalize the ρ'_k coordinates in order to have them in the range $0 \leq \rho'_k \leq 1$. This is accomplished by dividing each ρ'_k by $\rho'_{max} = \max(\rho'_k)$, with $k = 1..r$.

In this way we get the final signature

$$signature(S) = \left[\frac{\rho'_1}{\rho'_{max}} \dots \frac{\rho'_r}{\rho'_{max}} \right]' \quad (9)$$

This representation just lacks invariances to in-plane rotation and to mirror symmetry. Fortunately, these complex transformations in the image, translate to very simple transformations of the signature vector. Namely rotation in the plane perpendicular to the line that passes through the center of the camera sensor is mapped to a permutation P_r of the signature vector, up to orientation errors (due to the sampling from the silhouette) of $\frac{\pi}{r}$ radians. Mirror symmetry in that same plane is mapped to a permutation P_s .

In particular a rotation by an angle of $\frac{2\pi}{r}$ corresponds to the cyclic permutation:

$$P_r = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & \vdots & \vdots & \ddots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} \quad (10)$$

Rotations by $\frac{2n\pi}{r}$ map to P_r^n . For example, let $k = [k_1 \dots k_m]'$ be a signature vector. Then $P_r^n k = [k_n \dots k_m k_1 \dots k_{n-1}]'$.

The mirror symmetry across the line that passes through the centroid and the point whose magnitude is the first element of the signature vector is given by $P_s k$, with

$$P_s = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ \vdots & \dots & 0 & 1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 0 \end{bmatrix} \quad (11)$$

In this case $P_s k = [k_1 k_m k_{m-1} \dots k_2]$.

4 The Classifier

As the kinds of invariance we desire are mapped to two specific permutations of the feature vector, our problem gets easier than the general Linear Assignment Problem. In fact, we go from $m!$ different possible assignments in the general permutation case, to just $m \cdot 2$, with m corresponding to the rotations, and the 2 to the mirror symmetry. The most efficient way to solve the LAP problem under this constraints is the evaluation of all hypothesis, which is $O(n)$, while the Kuhn-Munkres algorithm is $O(n^3)$. The only change required to algorithm 1 is then step 3: we should instead evaluate all the $m \cdot 2$ valid transformations of a signature, and choose the resulting permutation for which the reward $\lambda y_i w^j x'_i + c^j x'_i$, under an appropriate norm, is largest. The resulting situation can be better understood by seeing it as having a matrix of transportation prizes, from N factories to N warehouses, with the constraint that once you assign one factory to a warehouse, only two scenarios remain possible, and in both all the assignments are uniquely determined. For example, consider the following reward matrix:

$$R = \begin{bmatrix} 1 & 3 & 2 & 1 \\ 6 & 2 & 5 & 2 \\ 0 & 1 & 1 & 3 \\ 2 & 4 & 2 & 1 \end{bmatrix} \quad (12)$$

The LAP problem refers to finding the set of four r_{ij} elements, with no i and no j repeated, whose sum is maximal. This sum can be seen as an l1 norm. The solution in this case is $[r_{21} r_{42} r_{13} r_{34}]$ with total reward 15. Using our restrictions on the possible assignments, the solution would be $[r_{21} r_{12} r_{43} r_{34}]$ with total reward 14.

In order to illustrate the effect of the algorithm on the input signatures, and the meaning of different values of λ , it's useful to observe figure 4. In (a) 3 patterns from 7 different classes are initially with different rotations and mirror symmetries. In (b) are the same patterns after being transformed with a high λ , and in (c) with low λ . The SVM was trained in a one-vs-all scheme, with the one being the class of the patterns in the first row. The effect of high λ was to "encourage" a higher margin between the first class and all the others, and that is easily visible: the postures in the first class are oriented in a different direction than the others. In (c) they are all aligned, they're enclosed by a smaller hypersphere.

5 Experimental Results

Due to the inexistence (to the best of our knowledge) of specialized image databases , we collected ourselves 50 samples from 7 different postures, with different scales, orientations and mirror symmetries. The samples represent hand gestures of 5 different adult male users, whose hand silhouettes were sampled to 80 points, after a skin color segmentation process. The postures considered are depicted in figure 5.

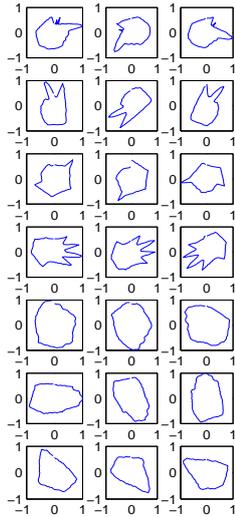
In order to evaluate the performance of the permutation invariant SVM on the data, we used it with a linear kernel, and compared against a normal SVM with a radial basis kernel applied on a regular feature vector, and on another feature vector which employed a popular heuristic to provide some invariance to rotation: selecting as the first value of the signature the one with largest magnitude and permuting cyclically the other elements of the signature accordingly. Using an SVM with this feature can, loosely speaking, be interpreted as an approximation to the permutation

σ	SVM	SVM+heuristic	pi-SVM($\lambda = 0.001$)
0.005	62%	94%	95%
0.007	56%	83%	94%
0.01	46%	47%	70%

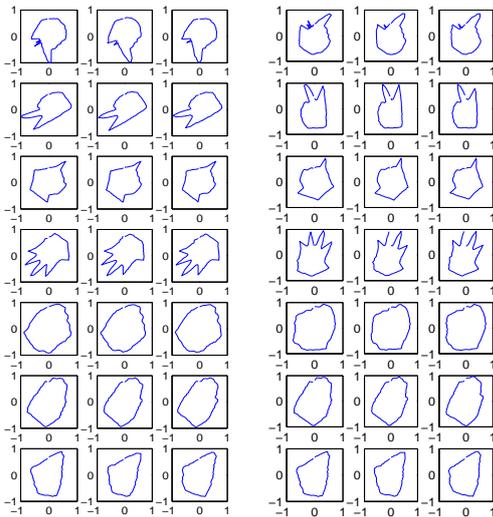
Table 1: Percentage of correct classifications with gaussian noise, with zero mean and standard deviation σ .

d	SVM	SVM+heuristic	pi-SVM($\lambda = 0.001$)
0.005	62%	86%	95%
0.01	56%	78%	88%
0.05	60%	68%	79%

Table 2: Percentage of correct classifications with salt & pepper noise (changes a d fraction of the signature points to magnitude 0 or 1).



(a)



(b)

(c)

Figure 4: The effect of λ in the resulting patterns. In a) are the input patterns. In b) they are permuted after training a SVM for the class in the first row against the others, with high λ . In c) after training with low λ .



Figure 5: Predefined set of postures.

invariant SVM with the restriction to permutations that rotate the feature, when solving the constrained LAP problem using the l_∞ norm. A C value of 2 was used in all experiments, so that we could focus on factors more directly connected to the algorithm under scrutiny.

We chose the number of iterations of the algorithm to be 4, because we observed that usually it was enough for convergence. The number of samples used in training was 10 from each class; the rest was used for testing. In order to test the robustness of the different solutions we applied two kinds of noise, gaussian and salt and pepper (also known as on-off), and averaged the results over 5 sessions. In general terms, gaussian noise changes all the components of the feature vector by small amounts, while salt and pepper turns a few components to zero or one. The results are shown in tables 1 and 2, and the aspect of noisy examples is shown in figure 6.

Finally, we also tried different norms for solving the LAP problem - table 3. The measure of quality used in all experiments was the percentage of correct classifications in the test set.

5.1 Discussion of the Results

One curious observation was that for large λ the permutation invariant SVM performed very poorly. This was only verified in the multi-class scenario. Preliminary tests on two class discrimination didn't show this phenomenon, quite the opposite. This may be explained by a poor fit of the one-versus-all way of combining binary classifiers, which we employed.

For small λ the method performed better than the other solutions, specially when the data was noisy. In these cases it greatly outperformed the other methods.

We used a permutation invariant SVM with a linear kernel, and we think that using a kernel that creates a non-linear decision function can improve the performance of the method - especially if we can't

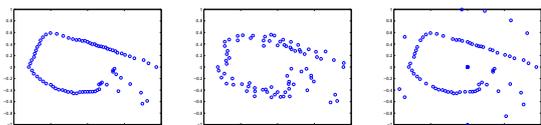


Figure 6: An example of the posture “two” is depicted on the left. In the center it's the same example, but with gaussian noise with mean 0 and standard deviation 0.005. On the right it's with salt and pepper noise, with 10 % of the components changed.

λ	pi-SVM(l_1)	pi-SVM(l_2)	pi-SVM(l_∞)
0.001	95%	99%	93%
0.1	90%	96%	88%
10	60%	74%	63%

Table 3: The effect of λ and the norm employed in solving the constrained LAP on the percentage of correct classifications.

solve the problem of transforming the data to have large margin - but this improvement comes with a performance price.

Of the norms employed in solving the constrained LAP problem, the one that behaved the best was the l_2 norm.

6 Conclusion

We presented a specialization of the permutation invariant SVM for classification of silhouette signature features. While the features used are too limited for general shape classification, because they are difficult to extract from images and cannot represent rich shapes - in particular those having important internal traits - they make a good fit for posture recognition over tabletops:

- They're cheap to compute, which is important in an input device (the mouse doesn't steal many cpu cycles).

- Certain interesting invariances appear as simple permutations of the feature vector, and this enables the use of the permutation invariant SVM efficiently.

- In tabletops powered by LCD displays, we can control the hand's background so that it is more easily segmentable.

The proposed method was shown to produce better results than other approaches, namely simple SVM classification with and without some common heuristics, using our data set of hand postures from five individuals. We have yet to evaluate the robustness of the approach to hand morphologies that are not in the database, like from children, but our preliminary results with synthetic noise looked quite promising.

Something that the proposed method apparently precludes is linear dimensionality reduction of the feature vectors (for example with PCA). The problem is that it's not possible to explore permutations of the features in a reduced linear space. We would have to transform the features back to the original space, perform the permutations and then transform the features back to the reduced dimensionality space. Maybe using nonlinear methods would work, like kernel PCA or spectral methods, but that would come with performance penalties.

Future work includes exploring different paradigms for combining the binary classifiers. One versus the rest doesn't appear to work well with the permutations of the data. A possibility is to experiment using the multi-class SVM (Weston 1999). We're also considering ways to extend the silhouette feature to include information about the internal traits of the shape.

References

Baraldi, S., Bimbo, A., Landucci, L. & Valli, A. (2006), witable: finger driven interaction for collaborative knowledge-building workspaces, in 'Computer Vision and Pattern Recognition Workshop, 2006 Conference on', pp. 144-144.

Belongie, S., Malik, J. & Puzicha, J. (2002), 'Shape matching and object recognition using shape contexts', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(4), 509-522.

Bi, J. & Zhang, T. (2004), Support vector classification with input data uncertainty, in 'Advances in Neural Information Processing Systems'.

Derpanis, K. (2004), 'A review of vision-based hand gestures'.

Dietz, P. & Leigh, D. (2003), 'Diamondtouch: a multi-user touch technology', *ACM Symposium on User Interface Software and Technology (UIST)* **1-58113-438-X**, 219-226.

Letessier, J. & Bérard, F. (2004), Visual tracking of bare fingers for interactive surfaces, in 'UIST '04: Proceedings of the 17th annual ACM symposium on User interface software and technology', ACM Press, New York, NY, USA, pp. 119-122.

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vision* **60**(2), 91-110.

Morris, M.R., H. A. P. A. & Winograd, T. (2006), Cooperative gestures: Multi-user gestural interactions for co-located groupware, in 'Proceedings of CHI', pp. 1201-1210.

Papadimitriou, C. H. & Steiglitz, K. (1982), *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall.

Sato, Y., Kobayashi, Y. & Koike, H. (2000), Fast tracking of hands and fingertips in infrared images for augmented desk interface.

Scholkopf, B. & Smola, A. J. (2002), *Learning with kernels: Support vector machines, regularization, optimization, and beyond.*, MIT Press.

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Shivaswamy, P. & Jebara, T. (2006), Permutation invariant svms, in 'International Conference on Machine Learning'.

Vapnik, V. (1995), *The nature of statistical learning theory*, Springer-Verlag.

Weston, J., W. C. (1999), Support vector machines for multi-class pattern recognition, in 'ESANN'.

Wu, M. & Balakrishnan, R. (2003), 'Multi-finger and whole hand gestural interaction techniques for multi-user tabletop displays', pp. 193-202.

Patch-Based Representation of Visual Speech

Patrick Lucey

Sridha Sridharan

Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
Brisbane, QLD, 4001, Australia

Abstract

Visual information from a speaker's mouth region is known to improve automatic speech recognition robustness, especially in the presence of acoustic noise. To date, the vast majority of work in this field has viewed these visual features in a holistic manner, which may not take into account the various changes that occur within articulation (process of changing the shape of the vocal tract using the articulators, i.e lips and jaw). Motivated by the work being conducted in fields of audio-visual automatic speech recognition (AVASR) and face recognition using *articulatory features* (AFs) and *patches* respectively, we present a proof of concept paper which represents the mouth region as an ensemble of image patches. Our experiments show that by dealing with the mouth region in this manner, we are able to extract more speech information from the visual domain. For the task of visual-only speaker-independent isolated digit recognition, we were able to improve the relative word error rate by more than 23% on the CUAVE audio-visual corpus.

Keywords: Visual Speech Recognition (VSR), Patches, Articulatory Features (AFs).

1 Introduction

Over the past twenty years, considerable research activity has concentrated on utilizing visual speech extracted from a speaker's face in conjunction with the acoustic signal, in order to improve robustness of automatic speech recognition (ASR) systems (Potamianos, Neti, Gravier, Garg & Senior 2003). Critical to the performance of the resulting audio-visual ASR (AVASR) system is the choice of visual features that contain sufficient information about the uttered speech (Potamianos & Scanlon 2005). Even though the visual features used over this time have shown to improve robustness to the overall AVASR system in extreme noisy conditions, the visual-only speech recognition (VSR) performance in these systems do lag by over an order of magnitude to its acoustic counterpart in clean conditions (Potamianos et al. 2003). This fact, clearly highlights the lack

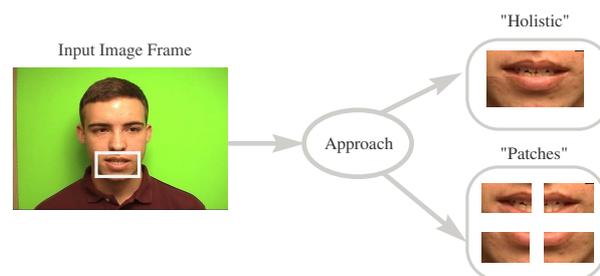


Figure 1: Following the extraction of the ROI, we propose to extract and model the ROI as an ensemble of image “patches” instead of the “holistic” approach which is currently being used in AVASR literature.

of speech classification power current visual features possess to extract speech information to the level of its acoustic counterpart. It may be the case that the visual modality does not hold as much information as the acoustic modality, however, this has not yet been quantified which motivates this research.

In AVASR literature, there have been numerous different methods of extracting visual features from the mouth *region of interest* (ROI) (see Section 2). However, all of these techniques modelled the ROI in a holistic, single stream manner. A potential problem which may arise from this approach is that these features may not take into account all of the various changes that occur within the mouth region during articulation (process of changing the shape of the vocal tract using the articulators, i.e lips and jaw) (Fant 1960). In contrast to the majority of work being conducted in the field of VSR, Saenko et al. has recently proposed the use of multiple streams of hidden *articulatory features* (AFs) to model the visual domain (Saenko, Darrel & Glass 2004). In this work, each sound is described by a unique combination of various articulator states, such as “lip-opened”, “lip-rounded”, “presence of teeth” etc.

Multi-stream approaches have also been used to good effect in the field of face recognition. Techniques that decompose the face into an ensemble of salient *patches* have reported superior face recognition performance with respect to approaches that treat the face as a whole (Brunelli & Poggio 1993, Moghadam & Pentland 1997, Martinez 2002, Kanade & Yamada 2003). The idea behind breaking the face into patches is that it is easier to take into account changes in appearance due to the faces complicated three-dimensional shape, in comparison to treating it holistically (Lucey & Chen 2006).

Heavily motivated by the work being conducted with patches in face recognition and AFs in AVASR, we present a novel approach to VSR by breaking the ROI into a series of image patches (see Figure 1).

This research was supported by the Australian Research Council Grant No: LP0562101

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

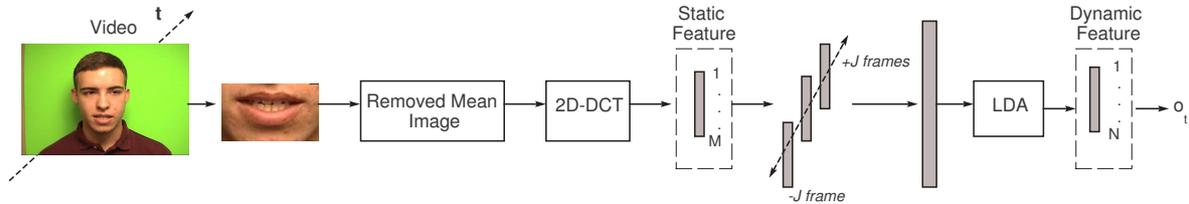


Figure 2: Block diagram of visual feature extraction process.

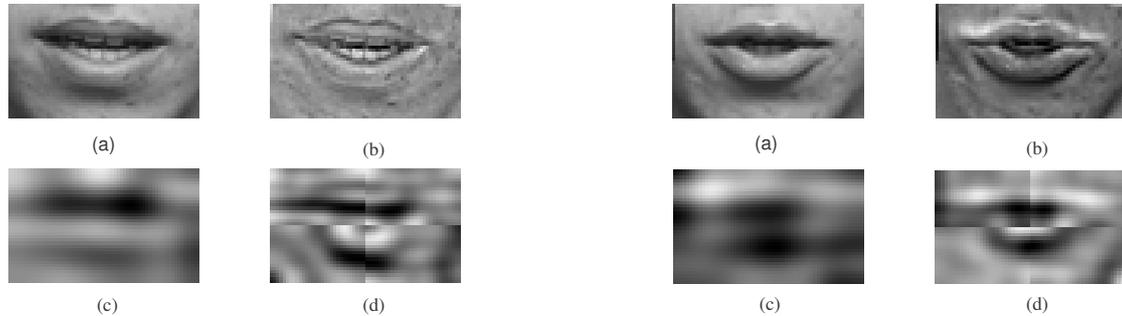


Figure 3: An example ROI from a speaker uttering the phoneme /th/ in the digit “three”. (a) original image, (b) mean-removed image, (c) reconstructed “holistic” image showing just the mouth somewhat opened, and (d) reconstructed “patch-based” image, displaying the presence of teeth and lip protrusion.

Figure 4: An example ROI from a speaker uttering the phoneme /uh/ in the digit “two”. (a) original image, (b) mean-removed image, (c) reconstructed “holistic” image showing just lip openness information, and (d) reconstructed “patch-based” image, displaying the presence of lip roundness and protrusion.

It is hoped by modelling each patch separately, we can take advantage of the local information contained within each patch, and also monitor any dynamic changes that occur during articulation.

By approaching visual speech in this manner, we hope to extract more speech information which will hopefully in turn increase the overall performance of VSR. A benefit of the following approach is that we are able to avoid the *curse of dimensionality* (Chatfield & Collins 1991) by alleviating the restriction of the number of visual features able to be used. This is our main motivation behind this work and is described and discussed in some detail in Section 2.

Following that, Section 3 describes the baseline VSR system, namely ROI detection and tracking, and the holistic visual feature extraction technique and modelling details. Section 4 describes the Patch-based VSR system. Section 5 presents our experimental results, and, finally Section 6 concludes the paper with a summary and a few remarks.

2 Motivation for Patch-Based Approach

Visual speech features can be categorized into two types, namely: area, and contour based representations. Area-based representations are concerned with transforming the whole *region of interest* (ROI) mouth pixel intensity image into a meaningful low-dimensional feature vector. Such transforms used for this approach include *principal component analysis* (PCA) (Bregler & Konig 1994), discrete cosine transform (DCT) (Heckmann, Kroschel, Savariaux & Berthommier 2002), linear discriminant analysis (Bellhumeur, Hespanha & Kriegman 1997) or a combination of DCT and LDA (Potamianos et al. 2003). Contour based representations, are concerned with parametrically atomising the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.) (Wark & Sridharan 1998). An *Active Appearance Model* (AAM) (Cootes, Edwards & Taylor 1998), combines

both the area and contour parameters together into a single feature vector. None of these above approaches have shown themselves to be clearly superior to each other, but due to its ability to be computed quickly, most researchers have preferred to use the area-based representation, as highlighted by the review conducted by Potamianos et al. (2003).

For area-based features, the current state-of-the-art consists of a hierarchical process. It is based on the hierarchical LDA (or *HiLDA*) process devised by Potamianos et al. (2003) and is shown in Figure 2. Firstly, the mouth ROI is extracted and features extracted using the two-dimensional DCT. The top M energy features are then selected to give a compact representation of the ROI. This resulting vector is called the *static feature*. This static feature vector is then concatenated with $\pm J$ adjacent frames and then LDA is used to project it down to N features giving the resultant *dynamic feature* vector o_t (See Section 3.2 for full description).

In literature, some researchers use only the top 20-30 DCT or PCA (very similar performance to DCT) coefficients for their static feature (Gowdy, Subramanya, Bartels & Bilmes 2004, Heckmann et al. 2002, Liang, Liu, Zhao, Pi & Nefian 2002). Potamianos et al. (2003) use the top 100 features, then use LDA to project it down to 30 features. As dynamic features provide the most information about speech (Goldschen, Garcia & Petajan 1994), it is necessary to keep the number of static features low, as computing the LDA matrix for high input features is computationally prohibitive (hence the reason why 20-30 static features are used). However, it is our contention that limiting the number of static features to around this number limits the amount of available speech stemming from the visual modality. This contention is backed up by the work conducted by Potamianos and Scanlon (Potamianos & Scanlon 2005), as they proposed another way of overcoming the dimensionality problem of the static feature vector. In this work, they made use of the laterally symmetric nature of

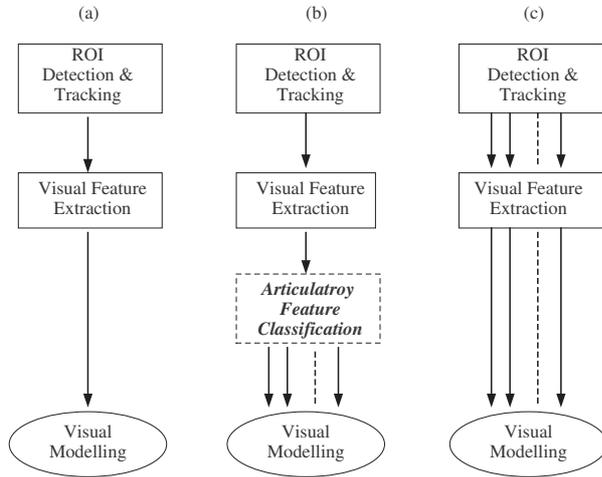


Figure 5: Comparison of the various approaches to visual speech recognition. (a) Shows the holistic approach currently being used in VSR. (b) Shows the multi-stream approach using articulatory features. (c) Shows our patch-based approach, which differs as each patch is treated independently from the initial ROI detection and tracking module.

a speaker’s lips by removing the odd frequency discrete cosine transform (DCT) components from the selected visual feature vector. By removing redundancies in the frequency domain, they reported some improvement in visual speech classification.

However, in an effort to get away from conventional holistic techniques and inspired by the work conducted in face recognition with patches, we sought motivation from the following examples shown in Figures 3 and 4. In VSR systems like our baseline one (see Section 3), initially the mean ROI image is subtracted to remove speaker dependencies (Figure 3b and 4b). Due to dimensionality restrictions, only the top 30 DCT coefficients are then extracted from each frame. Upon reconstruction of these images using the 30 top DCT coefficients, it can be seen that not much mouth information is visible (Figure 3c and 4c). Only maybe the mouth being open, and some coarse shape information is retained. However, when you view the original mean-removed images, it can be seen that other important visible articulatory information information such as the presence of teeth (Figure 3b) or lip roundness and protrusion (Figure 4b) is omitted. However, if we break the ROI images into patch quadrants, and use the top 30 DCT coefficients per patch, we are able to gain a closer representation of the original ROI, obviously due to the four-fold increase in features (Figure 3d and 4d). In Figure 3d, teeth information is present, along with lip protrusion and mouth opening information. In Figure 4d not only is it visible that the mouth is open, lip protrusion and roundness information can be seen.

Obviously by using more features, we are able to see more detail in the images. However, this example shows the benefit of using patches, as each patch can be modelled separately, hence overcoming the dimensionality restriction enforced on the static feature vector by the holistic single-stream topology. This approach is similar to Saenko et al. (2004), where they used multiple streams of hidden *articulatory features* (AFs) to model the visual domain. However, this approach requires additional complexity to the overall VSR framework, where each of these articulatory states (such as “lip-opened”) require extra classification (via a Support Vector Machine) prior to the

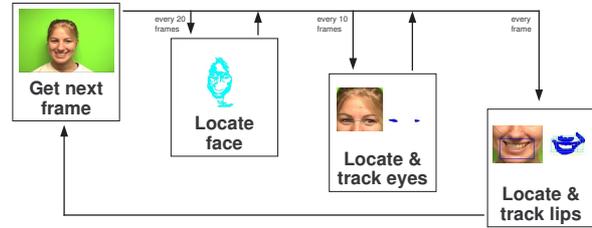


Figure 6: Overview of lip tracking system.

sound classification. The differences between all 3 approaches are shown in Figure 5.

In this paper, we show by representing the ROI as an ensemble of independent patches, we are able to obtain more visible speech information from the static features, in turn improving the overall visual speech recognition performance. This is shown through the improvement in performance for the task of speaker-independent isolated digit recognition on the CUAVE database (Patterson, Gurbuz, Tufekci & Gowdy 2002).

3 Baseline Visual Speech Recognition System

We now proceed to briefly components of our baseline VSR system. There exist three main components, which are over-viewed in the next three subsections: (a) visual front-end; (b) visual feature extraction; and (c) the visual modelling step. This baseline VSR system will be compared our patch-based system in Section 4.

3.1 Visual Front-End

Before the visual speech features can be extracted, the ROI has to be detected and tracked. In an AVASR system, this is performed by the visual front-end. For AVASR to be effective, it is essential that the visual front-end be highly accurate, otherwise these errors will cascade throughout the system and have a large effect on the ability of the final AVASR system to reliably recognize speech. This is known as the *front-end effect*.

In this study, the visual front-end consisted of three stages; face location, eye location and lip location. As shown in Figure 6, each stage was used to help form a search region for the next stage.

3.1.1 Face Location

Before face location was performed on the videos, 10 manually selected skin points for each speaker are used to form thresholds for the red, green and blue (r, g, b) values in colour-space for skin segmentation. The thresholds for each colour-space were calculated from the skin points as

$$\mu_c - \sigma_c \leq p_c \leq \mu_c + \sigma_c, \quad (1)$$

Where $c \in \{r, g, b\}$, μ_c and σ_c are the mean and standard deviation of the 10 points in colour-space c , and p_c is the value of the pixel being thresholded in colour-space c .

Once the thresholds were calculated, they were used for skin segmentation of the video to generate a bounding box of the face region within the frames every 20 frames, and this face location was remembered in the intermediate frames.

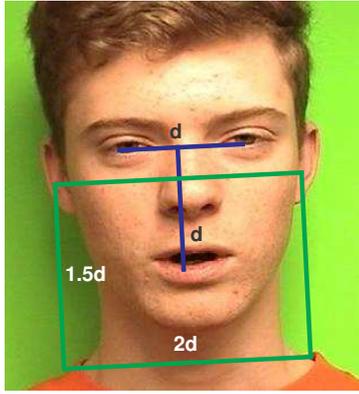


Figure 7: Calculating lip search region from eye locations.

3.1.2 Eye Location and Tracking

When transformed into $YCbCr$ space, the eye region of face images exhibit a high concentration of blue-chrominance, and a low concentration of red-chrominance. Therefore eye detection can be done in the $Cr - Cb$ space with reasonable results. However, eyebrows often appear as false positives and can degrade results. To remove the influence of eyebrows the $Cr - Cb$ image can be shifted vertically and subtracted from the original $Cr - Cb$ image. This will cancel the eyebrow minima by subtracting the eye minima, whereas the eye minima will be subtracted by the high values in the skin region and receive a large negative value suitable for thresholding (Butler, McCool, McKay, Lowther, Chandran & Sridharan 2003).

To locate the eyes from the face region from the previous stage, the top half of the face region was designated as the eye search-area, which was then searched using the shifted $Cr - Cb$ algorithm for the eye locations. The possible eye candidates were searched for two points that were not too large, too close horizontally, and not too distant vertically. Finally the two candidates which had the largest horizontal distance were chosen to be the eye locations. This process was performed every 10 frames, and the locations were remembered in the intermediate frames.

3.1.3 Lip Location and Tracking

Once the eye locations have been found, they are used to calculate a lip search region, as shown in Figure 7. The lip search region is then rotation-normalised, converted to R/G colour-space, and thresholded. The lip candidates from the thresholding are examined to remove unlikely lip locations (eg. too small, wrong shape). A search-window of 125×75 pixels is then scanned over the lip candidate image to find the windows with the highest concentration of lip candidate regions. The final lip ROI is chosen as the lowest, most central of these windows. Once the ROI was correctly located, the detected ROI was converted to grayscale and downsampled to 60×36 pixels for the experiments.

3.2 Visual Feature Extraction

The visual feature extraction process is given in Figure 2. Following the ROI extraction, the mean ROI over the utterance is removed. For purposes of notation the mouth ROI image matrix $I(x, y)$ is also expressed as the vectorised column vector $y = \text{vec}(I)$. So the mean removed mouth sub-image y^* is cal-

culated from a given temporal mouth sub-image sequence $Y = \{y_1, \dots, y_T\}$ such that,

$$y_t^* = y_t - \bar{y}, \text{ where } \bar{y} = \frac{1}{T} \sum_{t=1}^T y_t \quad (2)$$

This approach is very similar to cepstral mean subtraction used on acoustic cepstral features to improve recognition performance by providing some invariance to unwanted variations such as speaker dependencies. It is also similar to the *feature mean normalisation* of Potamianos et al. (2003), however in our approach we remove the redundant “DC” component in the image domain, instead of in the feature domain. A two-dimensional, separable, discrete cosine transform (DCT) is then applied to the resulting mean-removed image, with the $M = 30$ top DCT coefficients according to the zig-zag pattern retained, resulting in a “static” visual feature vector. Subsequently, to incorporate dynamic speech information, 21 neighboring such features over $\pm J = 10$ adjacent frames were concatenated, and were projected via an *inter-frame* LDA cascade to $N = 60$ dimensional “dynamic” visual feature vector.

3.3 The Speech Recognition System

In our experiments, we will be comparing two VSR systems: this baseline system, and our patch-based system (see Section 4). Both systems were designed to recognize isolated digits. As we are fusing multiple streams of data together, we saw isolated speech recognition as an ideal way to test our patch-based concept as it is easily implemented by calculating the likelihoods for the visual observations for a given word model. The continuous speech recognition paradigm is a much more complicated task as the number of possible hypothesis of word sequences becomes very large, and the number of best hypothesis obtained for each stream might not necessarily be the same. Our future work will concentrate on the continuous speech scenario, through the implementation of a Dynamic Bayesian Network (DBN) (Gowdy et al. 2004), which provides a framework to combine multiple streams together effectively.

In these experiments, each of the digits were modelled using 9 states and 18 Gaussians per state using HTK (Young, Everman, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev & Woodland 2002). These models were bootstrapped from the timed labelled transcriptions provided with the database. This topology was used as experimental and heuristic evidence showed that this was the optimal configuration.

4 Patch-Based Visual Speech Recognition System

The patch-based VSR system is very similar to that of the holistic baseline system, which was described in the previous section. The overall system is depicted in Figure 8. As it can be seen from the figure, this system is very basic. Essentially it is the baseline system being split into four parallel streams. The intended reason for this simple structure was to show that this configuration could be implemented easily. Also, by only breaking the ROI only into quadrants patches (no overlapping), we wanted to illustrate the benefit of treating parts of the ROI locally instead of as a whole.

As can be seen in Figure 8, the patch-based system uses the same visual front-end as the baseline system. Once the ROI has been detected and tracked, each grayscale 60×36 ROI image is broken up into $30 \times$

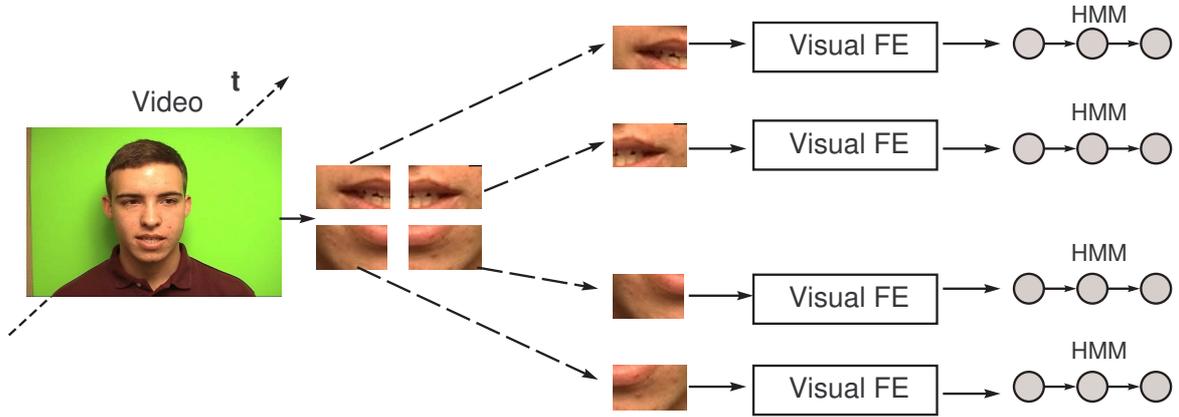


Figure 8: Block diagram of visual feature extraction process using the patch-based representation.

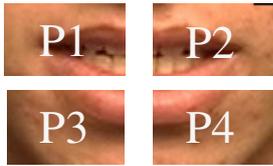


Figure 9: Once the mouth ROI has been detected and tracked, each ROI is broken up into quadrants and labelled.

18 quadrants (labelled as per Figure 9). Each one of these patches are then independently and visual features are extracted and modelled as per the process described in Section 3.2 and 3.3 respectively.

As mentioned previously, as a proof of concept we just conducted these experiments for the task of speaker-independent isolated digit recognition. As this was the case, fusion of the patches was performed via the weighted sum rule. Hence, let each spoken word be represented by a multiple visual speech observations \mathbf{O} , defined as

$$\mathbf{O} = \mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_R \quad (3)$$

where \mathbf{O}_r refers to the sequence of visual speech observations with regard to patch r . The isolated digit recognition can then be regarded as that of computing

$$\arg \max_{i=1}^{10} \left\{ \sum_{r=1}^R \beta_r P(\omega_i | \mathbf{O}_r) \right\} \quad (4)$$

where ω_i is the i 'th digit and β_r refers to the assigned patch weight. Also it is worth noting that $\sum_{r=1}^R \beta_r = 1$, where $0 > \beta_r > 1$.

5 Experimental Results

We now proceed to report a number of experimental results on the performance of the developed patch-based VSR system. The experiments were conducted on the CUAVE database.

5.1 The CUAVE Audio-Visual Corpus

For this work, we compared the speaker-independent visual-only isolated speech recognition performances on our baseline and patch-based systems. Training and evaluation visual speech was taken from the Clemson University, *CUAVE*, audio-visual database

(Patterson et al. 2002). The CUAVE database was selected as it is presently the only common audio-visual database which is available for all universities to use. This is important for benchmarking and comparison purposes. The CUAVE database consists of two major sections, one of individual speakers and one of speakers pairs. For this study, only the stationary connected-digit string section of the individual speakers were used. The stationary connected-digit string section of the database consisted of each of the 36 individual speakers uttering the connected digits “zero” to “nine” a total of 5 times each. The 36 individual speakers were divided arbitrarily into a set of 28 training speakers and 8 different test talkers for a completely speaker-independent grouping. As the database is so small, we used 10 different permutations of this configuration to see the effect of having different speakers in the training/testing set.

5.2 Isolated Digit Recognition Results

Generally, an accurate measure of how much speech information is contained within the visual features is indicative of how well it performs in the task it is being used for, which in this case is isolated digit VSR. We first performed this on the *static* visual features for the holistic (H), patch-based (P), fused holistic and patch-based features (F), patches concatenated (PC), and patches and holistic concatenated (FC). The first experiment was conducted using the same amount of features as the holistic system (i.e. $M = 30$ for H, P, F, PC and FC). For P, 8 features were used for $P1$ and $P2$ and 7 features for $P3$ and $P4$, and each patch was weighted equally. For F, 6 features were used for each patch quadrant and the holistic patch. For this configuration, the holistic approach was weighted 50% and each patch was weighted 12.5%. The PC and FC experiments were conducted to see the effect of modelling each patch independently instead of in a single stream.

The second experiment was conducted using the same method, however, the same amount of features were used for the patched-based system (i.e. $M = 120$ for H, P, F, PC and FC). For P, 30 features were used for $P1 - P4$. For F, 24 features were used for each patch quadrant and the holistic patch. The experiments were carried out in this way so that we could evaluate how much speech information there is for the same amount of features. The results are given in Table 1.

As can be seen in Table 1, using the same amount of features, the patch-based system outperforms the holistic system using both 30 and 120 features. And

Exp	H	P	F	PC	FC
1	57.10	44.72	44.27	66.25	55.16
2	58.69	45.38	44.80	63.73	56.17

Table 1: Isolated WERs of the static features for the: (H) holistic or baseline system, (P) patch-based system, (F) fused holistic and patch-based system, (PC) patches concatenated, (FC) holistic and patches concatenated. For experiment 1, $M = 30$ and for experiment 2, $M = 120$.

Exp	H	P	F
1	30.10	25.95	22.92
2	-	28.22	23.68

Table 2: Isolated WERs of the dynamic features concatenating ± 10 frames then using LDA to yield 60 features from the static features given in Table 1.

when the holistic and patch-based system were fused together more improvement was gained. It is somewhat interesting to note that the better performance was gained in experiment 1, and not 2, with the fused holistic and patch-based system achieving the best performance with a word-error-rate (WER) of 44.27% compared to 57.10% for the holistic system. This goes against our initial hypothesis regarding dimensionality, as lower number of features actually obtained around the same or marginally more static speech information. However, it may be the case that the top 30 features contain most of the speech information, whilst the remaining features contain mostly unique speaker information. Another interesting result is that modelling each patch independently seems to achieve better results than concatenating the features and modelling them as one (PC, FC). This may suggest that representation of features is the key to VSR, and not just the sheer number of features used. However, it must be noted that these results may not be significant due to the small size of the database and further investigation is need before any claims can be made about performance.

To gauge the overall performance of the systems using the full system (i.e incorporating the dynamic features); the holistic, patch-based, and fused holistic and patch-based system were compared. The results are given in Table 2. As can be seen from these results, the fused system was again was the best performed following the trend of the previous experiments. For experiment 1, the WER of 22.92% was much better than the holistic one of 30.10%, giving a 23.9% relative improvement. Again these results look very promising, but further investigation really needs to be done before determining whether these results are significant or not. It is also worth noting that no holistic result for experiment 2 could be gain as the dimensionality for the LDA matrix was too large to be computed.

6 Summary and Conclusion

In this paper, we presented a novel patch-based approach to the task of VSR which showed improvement over holistic approaches. Our results show that our concept of breaking up the mouth ROI into patches, instead of just one whole, could extract more speech information from the visual domain. We understand that a major limitation of our experiments was the small size of our training and testing database. How-

ever, we believe that the results give an indication that this patch-base approach is worth pursuing on an larger database, as well as on the more complicated task of continuous speech recognition. Our future work will concentrate on the continuous speech recognition scenario, through the implementation of a Dynamic Bayesian Network (DBN), which provides a framework to combine multiple streams together effectively. We believe the DBN framework is a far more prudent way to go rather than using feature fusion as this approach really is not practical as it does not allow us to weight the various patches and may cause catastrophic fusion. Another task we will be undertaking in the future will be investigating which patches in the ROI (or even the face) are most pertinent for visual speech (such as corner of mouths, mouth center, cheeks etc), so as to further enhance VSR.

7 Acknowledgements

We would also like to thank Clemson University for freely supplying us their CUAVE audio-visual database for our research.

References

- Belhumeur, P., Hespanha, J. & Kriegman, D. (1997), 'Eigenfaces vs Fisherfaces: Recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Machine Intell.* **19**(7), 711–720.
- Bregler, C. & Konig, Y. (1994), Eigenlips for robust speech recognition, in 'International Conference on Acoustics, Speech and Signal Processing', Vol. 2, Adelaide, Australia, pp. 669–672.
- Brunelli, R. & Poggio, T. (1993), 'Face recognition: Features versus templates', *IEEE Trans. PAMI* **15**(10), 1042–1052.
- Butler, D., McCool, C., McKay, M., Lowther, S., Chandran, V. & Sridharan, S. (2003), Robust face localisation using motion, colour and fusion, in C. Sun, H. Talbot, S. Ourselin & T. Adriaansen, eds, 'Seventh International Conference on Digital Image Computing: Techniques and Applications', CSIRO Publishing, Macquarie University, Sydney, Australia.
- Chatfield, C. & Collins, A. J. (1991), *Introduction to Multivariate Analysis*, London, United Kingdom: Chapman and Hall.
- Cootes, T., Edwards, G. & Taylor, C. (1998), Active appearance models, in 'Proc. Europ. Conf. Computer Vision', Germany, pp. 484–498.
- Fant, G. (1960), Acoustic theory of speech production.
- Goldschen, A. J., Garcia, O. N. & Petajan, E. (1994), Continuous optical automatic speech recognition by lipreading, in A. Singh, ed., 'Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers', Vol. vol.1, IEEE Comput. Soc. Press, Pacific Grove, CA, USA, pp. 572–577.
- Gowdy, J. N., Subramanya, A., Bartels, C. & Bilmes, J. (2004), DBN Based Mult-Stream Models for Audio-Visual Speech Recognition, in 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', Vol. 1, pp. 993–996.

- Heckmann, M., Kroschel, K., Savariaux, C. & Berthommier, F. (2002), Dct-based video features for audiovisual speech, *in* 'Proc. Int. Conf. Spoken Language Processing', pp. 1925–1928.
- Kanade, T. & Yamada, A. (2003), 'Multi-subregion based probabilistic approach towards pose-invariant face recognition', *IEEE International Symposium on Computational Intelligence in Robotics Automation* **2**, 954–959.
- Liang, L., Liu, X., Zhao, Y., Pi, X. & Nefian, A. (2002), Speaker Independent Audio-Visual Continuous Speech Recognition, *in* 'Proc. Int. Conf. on Multimedia and Expo', Vol. 2, pp. 25–28.
- Lucey, S. & Chen, T. (2006), Learning patch dependencies for improved pose mismatched face verification, *in* 'IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)'.
- Martinez, A. M. (2002), 'Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class', *IEEE Trans. PAMI* **24**(6), 748–763.
- Moghaddam, B. & Pentland, A. (1997), 'Probabilistic visual learning for object recognition', *IEEE Trans. PAMI* **19**(7), 696–710.
- Patterson, E. K., Gurbuz, S., Tufekci, Z. & Gowdy, J. N. (2002), CUAVE: a new audio-visual database for multimodal human-computer interface research, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', Orlando.
- Potamianos, G., Neti, C., Gravier, G., Garg, A. & Senior, A. W. (2003), 'Recent advances in the automatic recognition of audio-visual speech', *Proc. of the IEEE* **91**(9).
- Potamianos, G. & Scanlon, P. (2005), Exploiting lower face symmetry in appearance-based automatic speechreading, *in* 'Proceedings of the Auditory-Visual Speech Processing International Conference 2005', British Columbia, Canada, pp. 79–84.
- Saenko, K., Darrel, T. & Glass, J. (2004), Articulatory features for robust visual speech recognition, *in* 'Int. Conf. Multitmodal Interfaces'.
- Wark, T. & Sridharan, S. (1998), An approach to statistical lip modelling for speaker identification via chromatic feature extraction, *in* 'International Conference on Pattern Recognition', pp. 123–125.
- Young, S., Everman, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2002), *The HTK Book (for HTK Version 3.2.1)*, Entropic Ltd.

Audio-Visual Speaker Verification using Continuous Fused HMMs

David Dean¹Sridha Sridharan¹Tim Wark^{1,2}¹Speech, Audio, Image and Video Research Laboratory, Queensland University of Technology²CSIRO ICT Centre
Brisbane, Australia

ddean@ieee.org, s.sridharan@qut.edu.au, tim.wark@csiro.au

Abstract

This paper examines audio-visual speaker verification using a novel adaptation of fused hidden Markov models, in comparison to output fusion of individual classifiers in the audio and video modalities. A comparison of both hidden Markov model (HMM) and Gaussian mixture model (GMM) classifiers in both modalities under output fusion shows that the choice of audio classifier is more important than video. Although temporal information allows a HMM to outperform a GMM individually in video, this temporal information does not carry through to output fusion with an audio classifier, where the difference between the two video classifiers is minor. An adaptation of fused hidden Markov models, designed to be more robust to within-speaker variation, is used to show that the temporal relationship between video observations and audio states can be harnessed to reduce errors in audio-visual speaker verification when compared to output fusion.

Keywords: audio-visual speaker recognition (AVSPR), fused hidden Markov model (FHMM)

1 Introduction

The aim of audio-visual speaker recognition (AVSPR) is to make use of complementary information between the acoustic and visual domains to improve the performance of traditional acoustic speaker recognition. Most current approaches to AVSPR either combine the output of individual hidden Markov models (HMMs) in each modality (late fusion), or use a single HMM to classify both modalities (early fusion). Because the scores are combined at the whole-utterance level, late fusion cannot take true advantage of the temporal dependencies between the two modalities. While early fusion has the advantage that it can take advantage of these dependencies, it often suffers from problems with noise, and has difficulties in modeling the asynchronicity of audio-visual speech (Chibelushi, Deravi & Mason 2002). The problems with performing AVSPR with early or late fusion have led to the development of middle-fusion methods, or mod-

els that accept two streams of input and combine the streams *within* the model to produce a single score.

Most existing approaches to middle-fusion use coupled HMMs (Nefian, Liang, Fu & Liu 2003), which combine two single-stream HMMs by linking the dependencies of their hidden states. However, due to the small number of hidden states in each modality, these dependencies are often not strong enough to capture the true relationship between the two streams (Brand 1999). Fused HMMs (FHMMs) were developed (Pan, Levinson, Huang & Liang 2004) by attempting to design a model that maximises the mutual information between the two modalities within a multi-stream HMM. Pan et al. (2004) found that the optimal multi-stream HMM design would result from linking the hidden states of one HMM to the observations of the other, rather than linking the hidden states together, as in a coupled HMM.

In this paper, we first introduce a novel adaptation of Pan et al's FHMMs, designed to be more robust to within-speaker variation. A comparison of a number of different audio-visual output-fusion configurations is performed to obtain an insight into the temporal information available in both audio and video, individually and combined for the purposes of speaker verification. Finally we examine the ability of our FHMM model to take better advantage of the temporal dependencies between the modalities than is possible with output fusion alone.

2 Fused Hidden Markov Models

2.1 Theory

Consider two tightly coupled time series $\mathbf{O}^A = \{\mathbf{o}_0^A, \mathbf{o}_1^A, \dots, \mathbf{o}_{T-1}^A\}$ and $\mathbf{O}^V = \{\mathbf{o}_0^V, \mathbf{o}_1^V, \dots, \mathbf{o}_{T-1}^V\}$, corresponding to audio and video observations respectively. Assume that \mathbf{O}^A and \mathbf{O}^V can be modeled by two HMMs with hidden states $U^x = \{u_0^x, u_1^x, \dots, u_{T-1}^x\}$, where x is A or V , respectively. In the FHMM framework, an optimal solution for $p(\mathbf{O}^A; \mathbf{O}^V)$ according to the maximum entropy principle (Pan, Liang & Huang 2001) is given by

$$\tilde{p}(\mathbf{O}^A; \mathbf{O}^V) = p(\mathbf{O}^A) p(\mathbf{O}^V) \frac{p(\mathbf{w}, \mathbf{v})}{p(\mathbf{w}) p(\mathbf{v})} \quad (1)$$

where $\mathbf{w} = g_A(\mathbf{O}^A)$, and $\mathbf{v} = g_V(\mathbf{O}^V)$ are transformations designed such that $p(\mathbf{w}, \mathbf{v})$ is easier to calculate than $p(\mathbf{O}^A, \mathbf{O}^V)$, but still reflects the statistical dependence between the two streams. The final term in (1) can therefore be viewed as a correlation weighting, which will be high if \mathbf{w} and \mathbf{v} are related, and low if they are mostly independent. Pan et al. (2001) also showed that the minimum distance between $\tilde{p}(\mathbf{O}^A; \mathbf{O}^V)$ and the ground truth $p(\mathbf{O}^A, \mathbf{O}^V)$

This research was supported by a grant from the Australian Research Council (ARC) Linkage Project LP0562101.

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

is established when the mutual information between \mathbf{w} and \mathbf{v} is maximised:

$$(\hat{\mathbf{w}}, \hat{\mathbf{v}}) = \arg \max_{(\mathbf{w}, \mathbf{v}) \in \theta} \mathcal{I}(\mathbf{w}, \mathbf{v}) \quad (2)$$

In their audio-visual FHMM paper, Pan et al. (2004) chose \mathbf{w} and \mathbf{v} empirically from the following set (θ):

$$\mathbf{w} = \hat{\mathbf{U}}^A, \quad \mathbf{v} = \mathbf{O}^V \quad (3)$$

$$\mathbf{w} = \hat{\mathbf{U}}^A, \quad \mathbf{v} = \hat{\mathbf{U}}^V \quad (4)$$

$$\mathbf{w} = \mathbf{O}^A, \quad \mathbf{v} = \hat{\mathbf{U}}^V \quad (5)$$

where $\hat{\mathbf{U}}^x$ is an estimate of the optimal state sequence of HMM x for output \mathbf{O}^x . By invoking (2) over the set θ and invoking the following inequality in information theory

$$\mathcal{I}(x, f(y)) \leq \mathcal{I}(x, y) \quad (6)$$

And that estimated hidden state sequences can be viewed as a function of the output ($\hat{\mathbf{U}}^x = f_x(\mathbf{O}^x)$), Pan et al. (2004) concluded that

$$\mathcal{I}(\hat{\mathbf{U}}^A, \hat{\mathbf{U}}^V) = \mathcal{I}(\hat{\mathbf{U}}^A, f_V(\mathbf{O}^V)) \leq \mathcal{I}(\hat{\mathbf{U}}^A, \mathbf{O}^V) \quad (7)$$

$$\mathcal{I}(\hat{\mathbf{U}}^A, \hat{\mathbf{U}}^V) = \mathcal{I}(f_A(\mathbf{O}^A), \hat{\mathbf{U}}^V) \leq \mathcal{I}(\mathbf{O}^A, \hat{\mathbf{U}}^V) \quad (8)$$

Therefore the transforms (3) and (5) produce better estimates of $\tilde{p}(\mathbf{O}^A; \mathbf{O}^V)$ than (4). By invoking (3) in $p(\mathbf{O}^A; \mathbf{O}^V)$:

$$\begin{aligned} p_A(\mathbf{O}^A; \mathbf{O}^V) &= p(\mathbf{O}^A) p(\mathbf{O}^V) \frac{p(\hat{\mathbf{U}}^A, \mathbf{O}^V)}{p(\hat{\mathbf{U}}^A) p(\mathbf{O}^V)} \\ &= p(\mathbf{O}^A) p(\mathbf{O}^V | \hat{\mathbf{U}}^A) \end{aligned} \quad (9)$$

where $p(\mathbf{O}^A)$ can be obtained from the regular audio HMM and $p(\mathbf{O}^V | \hat{\mathbf{U}}^A)$ is the likelihood of getting the video output sequence given the estimated audio HMM state sequence which produced \mathbf{O}^A . This equation represents the *audio-biased* FHMM as the main decoding process is the audio HMM.

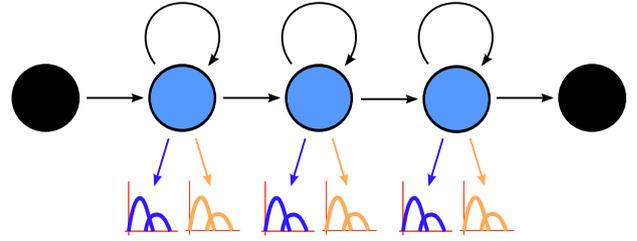
Similarly, invoking (5) to arrive at the *video-biased* FHMM gives:

$$p_V(\mathbf{O}^A; \mathbf{O}^V) = p(\mathbf{O}^V) p(\mathbf{O}^A | \hat{\mathbf{U}}^V) \quad (10)$$

The choice of the audio- or video-biased FHMM should be chosen upon which individual HMM can more reliably estimate the hidden state sequence for a particular application. Alternatively, both versions can be used concurrently and combined using output fusion, as in Pan et al. (2004).

2.2 Continuous FHMMs

In the original implementation of FHMMs (Pan et al. 2004), the subordinate modality features were treated as discrete symbols through vector-quantisation codebooks to simplify the calculation of the coupling parameters. However this simplification caused problems with within-speaker session variability, especially when the video was the subordinate modality. While audio-biased FHMMs (A-FHMMs) worked well in experiments on the CUAVE database (Dean,



(a) Fused HMM

Figure 1: State diagram representation of a FHMM. (Compare to a regular HMM in figure 2.)

		Configuration											
		1	2	3	4	5	6	7	8	9	10	11	12
Session	1	Train	Train	Train	Train	Eval	Test	Eval	Test	Eval	Test	Eval	Test
	2	Train	Eval	Test	Eval	Test	Train	Train	Train	Train	Test	Eval	Eval
	3	Eval	Test	Train	Test	Eval	Train	Test	Eval	Train	Test	Eval	Train
	4	Test	Eval	Test	Eval	Train	Test	Eval	Train	Train	Train	Train	Train

Table 1: XM2VTS dataset configurations used in these experiments

Wark & Sridharan 2006), the change in codebook values caused by a change in session outweighed that due to a change in speaker, rendering the discrete FHMM worse than the underlying HMM when used in a multi-session database like XM2VTS.

To allow the FHMM structure to more robustly model the subordinate modality, we proposed modeling the relationship between the dominant states and the subordinate observations using an extra GMM within each of the dominant states. Therefore our *continuous* FHMM (as opposed to Pan et al's *discrete* FHMM) can be viewed as a regular HMM with two GMM-based output probability distributions instead of one in a normal HMM, as shown in Figure 1.

3 Experimental Setup

3.1 Training and Testing Datasets

For this experiment, training, testing and evaluation data were extracted from the digit-video sections of the XM2VTS database (Messer, Matas, Kittler, Luetten & Maitre 1999). The training and testing configurations used for these experiments was based on the XM2VTSDB protocol (Luetten & Maitre 1998), but adapted to allow more tests than provided by the protocol. Each of the 295 speakers in the database has four separate sessions of video where the speaker speaks two sequences of two sentences of ten digits. In each of the configurations, two sessions were used for training, one for evaluation and one for testing, allowing for 12 configurations in total, as shown in Table 1. By comparison, the XM2VTSDB protocol only allows for the first configuration.

These experiments were performed as verification experiments, where the speaker would attempt to enter the system by claiming the identity of a particular client. To perform this task, the speakers were split into two groups: clients, who claimed their own identity; and imposters, who claimed the identity of one of the clients.

As per the XM2VTSDB protocol, 200 speakers were designated clients, and 95 were used as impostors. For each client testing sequence (2 per session), 20 sequences were chosen at random from the impostor set allowing for a total of 400 (200×2) client tests and 8000 ($200 \times 2 \times 20$) impostor tests for each configuration. Over all 12 configurations, 4800 client tests and 96000 impostor tests are performed.

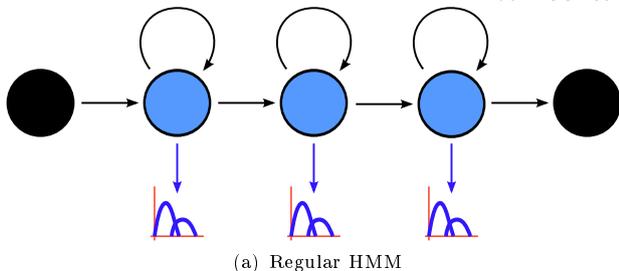


Figure 2: Regular HMM. The output probability of each state is implemented as a GMM.

3.2 Feature Extraction

Mel frequency cepstral coefficients (MFCCs) were used to represent the acoustic features in these experiments because of their general application to both speech and speaker recognition. Each feature vector consisted of the first 12 MFCCs, normalised energy coefficient, and the first and second time derivatives of those 13 features to result in a 43 dimensional feature vector. These features were calculated every 10 milliseconds using 25 millisecond Hamming-windowed speech signals.

Visual features were extracted from a manually tracked lip region-of-interest (ROI) from 25 fps (40 milliseconds / frame) video data. Manual tracking of the locations of the eyes and lips were performed every 50 frames, and the remainder of the frames were interpolated from the manual tracking. The eye locations were used to normalise the rotation of the lips. A rectangular region-of-interest, 120 pixels wide and 80 pixels tall, centered around the lips was extracted from each frame in the video. Each ROI was then reduced to 20% of its original size (24×16 pixels) and converted to grayscale. Finally the ROI was reduced to 20 dimensions using discrete cosine transformation (DCT) (Heckmann, Kroschel, Savariaux & Berthommier 2002). First and second time derivatives of these features were added to form a 60 dimensional feature vector.

4 Audio-Visual Speaker Verification using Output Fusion

4.1 Training

Two classifier-types were used for each modality, for a total of four output-fusion experiments. The two classifiers used were Gaussian mixture models (GMMs), which are good at modeling static, or time-independent, variables, and HMMs, which are better at modeling temporal variables. This can be observed by examining a standard HMM design: HMMs are commonly implemented as a chain of GMMs, as shown in Figure 2, where the HMM controls the likelihood of moving between states, and the GMM-states control the likelihood of outputting certain features when in an emitting state. Conversely, a GMM can be viewed as HMM with only one emitting state.

Both HMM and GMM speaker-dependent models were generated by adapting background models to each individual speaker. The background models were generated using the training sequences for each configuration over both clients and impostors. These models were then adapted to each individual client speaker’s training sequences using maximum a posteriori (MAP) adaptation (Lee & Gauvain 1993).

GMM models were trained over all training sequences, whereas HMM models were trained for each word. Empirical experiments were performed on

Model	Mixtures	States
Audio HMM	9	7
Audio GMM	256	-
Video HMM	16	7
Video GMM	8	-

Table 2: Best performing topologies for each classifier.

a single configuration to determine the best topology, shown in Table 2. HMM training was performed using the HTK toolkit (Young, Evermann, Kershaw, Moore, Odell, Ollason, Povey, Valtchev & Woodland 2002), and GMM training with internally developed utilities.

4.2 Testing

For each of the four client models trained in the previous section, two client sequences and 40 impostor sequences were verified using that model for each configuration. Scores obtained from the client models were normalised for length and environment by subtracting the background-model score for the same sequence.

In addition to the individual models, the four possible output-fusion combinations of audio and video classifiers were also examined, as listed below:

- Audio HMM + Video HMM
- Audio HMM + Video GMM
- Audio GMM + Video HMM
- Audio GMM + Video GMM

Given that the parameters of the score-distribution vary considerably between classifiers, the evaluation session of each configuration is used to get an estimation of each classifier’s score distribution, which is used to normalise the scores.

$$Z_i(s_i) = \frac{s_i - \hat{\mu}_i}{\hat{\sigma}_i} \quad (11)$$

Where s_i is the score from classifier i and $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the estimated mean and standard deviations of classifier i ’s score distribution. Therefore the output-fusion score for each combination is calculated as

$$s_F = \frac{Z_a(s_a) + Z_v(s_v)}{2} \quad (12)$$

Where a is the audio classifier and v is the video classifier.

4.3 Results

Detection error trade-off (DET) plots showing the performance of both the individual classifiers and the four output-fusion combinations for speaker verification are shown in Figure 3.

From a comparison of the HMM and GMM performance for each modality, it can be clearly seen that there is temporal information in both the audio and video features. Whilst the audio GMM performs nearly as well as the audio HMM, it is only through using a much higher number of mixtures (256 vs 9). However, in the video we found that the GMM performance could not be made to match the HMM’s, regardless of the number of mixtures used.

However, the clear improvement of using a video HMM over a video GMM does not appear to translate over to output fusion. The main differences in output fusion appears to be related to the audio classifier chosen and not the video. The video HMM does appear to improve output fusion slightly in areas of

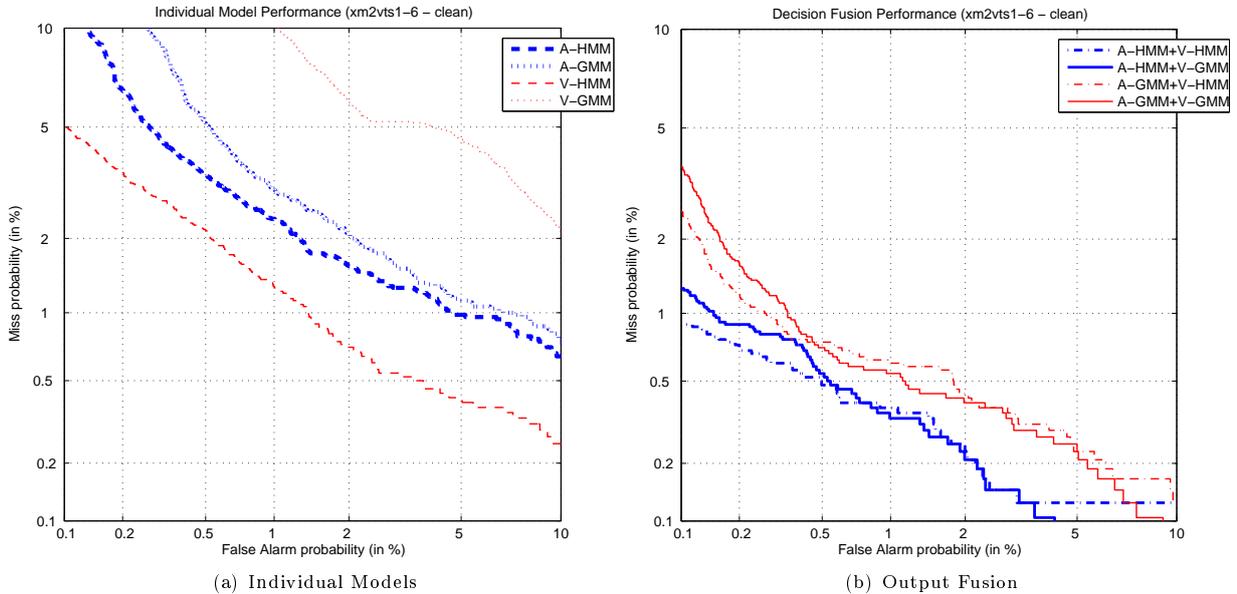


Figure 3: Detection error trade-off (DET) plots for output-fusion speaker verification.

low false alarm, but it does not provide a major improvement that the difference of the two classifiers in video alone might indicate. So, while the video HMM clearly takes advantage of temporal video information when compared to the video GMM, this temporal information provides little benefit in output fusion where a static GMM would work almost as well. It is also clear that output fusion cannot take advantage of temporal dependencies between the two modalities, as the only information fused together is the classifier’s scores over an entire utterance.

5 Audio-Visual Speaker Verification using FHMMs

5.1 Training

The training of a biased FHMM is a three-step process:

1. The dominant individual HMM is trained independently
2. The best hidden state sequence of the trained HMM is found for each training observation using the Viterbi process (Young et al. 2002)
3. The relationship between the hidden state sequences and the subordinate observations are modeled

For these experiments, both audio- and video-biased FHMMs were examined, so the underlying HMMs trained in Step 1 were the audio HMM and the video HMM as trained in Section 4.1, respectively.

The relationship between the hidden state sequences and the subordinate observations is contained in $p(\mathbf{O}^s | \hat{\mathbf{U}}^d)$ where d represents the dominant modality, and s the subordinate. This is basically defined as the likelihood of getting a subordinate observation when in a particular dominant state. Once the estimated hidden state sequence, $\hat{\mathbf{U}}^d$, for the training data was determined in Step 2, the subordinate training observations were segmented based on the word and state boundaries. Each speaker’s GMM (trained in Section 4.1) was then adapted for each word and state within their training sequences to form

the FHMM’s subordinate GMMs. The background GMM was also adapted to each word and state and added to the background HMM to form the background FHMM. The optimal number of mixtures for the subordinate GMMs was found empirically to be the same as that for individual GMM classifiers, that being 256 for the audio and 8 for the video.

The training sequence was performed twice, once with audio as the dominant modality, and once with video dominant to form the audio- and video-biased FHMMs respectively.

5.2 Testing

Generalising (9) and (10) we can see that:

$$p_d(\mathbf{O}^d, \mathbf{O}^s) = p(\mathbf{O}^d) p(\mathbf{O}^s | \hat{\mathbf{U}}^d) \quad (13)$$

Where d represents the dominant modality, and s the subordinate. As $p(\mathbf{O}^d) = \sum_{\mathbf{U}^d} p(\mathbf{O}^d, \mathbf{U}^d)$, and the aim of the decoding process is to find the optimal \mathbf{U}^d by maximising the likelihood, we find the optimal state sequence is given by:

$$\hat{\mathbf{U}}^d = \arg \max_{\mathbf{U}^d} p(\mathbf{O}^d, \mathbf{U}^d) p(\mathbf{O}^s | \mathbf{U}^d) \quad (14)$$

This can be viewed a special type of HMM that has two observation-emission probability-density-functions for each state, one being the continuous dominant-observation-emission GMM of the regular HMM, and the second being the continuous subordinate-observation-emission GMM trained in Section 5.1. As these scores are combined within each state, and each state still provides a single probability within the Viterbi process, the decoding process is otherwise unaffected.

Before the scores for each modality are combined within the state, they are normalised for each modality based on the evaluation data set, similar to the normalisation performed for output fusion in Section 4.2, but on a frame-by-frame basis rather than over an entire sequence. Because we found the difference in frame-scores between modalities is more significant that the difference in scores between speakers, the background dominant HMM and subordinate GMM individual models were evaluated for each

frame over the evaluation sequences for each configuration to come up with an estimate of each classifier's score distribution which was then used to normalise the GMM scores within each FHMM state using (12). The features evaluated for each modality's score is determined by the frame-rate of the dominant HMM, with the subordinate features chosen being the closest in time to the dominant features.

In addition to using models adapted to a specific word-state for the subordinate modality, models adapted to all states of a particular word, and just using the global speaker GMM in this role was considered. These three choices will be referred to as word-state GMMs, word GMMs and global GMMs for the remainder of this paper. By examining the difference in performance between these subordinate models in the FHMM structure, we can make some conclusions about the temporal dependencies captured by the FHMMs.

Finally, scores obtained from the client FHMM models were normalised for length and environment by subtracting the background-model FHMM score for the same sequence.

5.3 Comparison with Output Fusion

It can be seen that using the global speaker GMM should be functionally equivalent to a output fusion of the GMM and the underlying HMM. This is because at a base level the output HMM likelihood can be mathematically defined as:

$$p(\mathbf{O}) = \prod_t p_h(o_t|u_t) \quad (15)$$

Where $p_h(o_t|u_t)$ is the likelihood of the HMM outputting observation o_t whilst in state u_t at time t . Fusing the output of this HMM with a single GMM's output ($p_g(o_t)$) results in:

$$p(\mathbf{O}^d, \mathbf{O}^s) = \prod_t p_h(o_t^d|u_t^d) \times \prod_t p_g(o_t^s) \quad (16)$$

$$= \prod_t [p_h(o_t^d|u_t^d) p_g(o_t^s)] \quad (17)$$

This is equivalent to multiplying the regular HMM and global subordinate GMM within the Viterbi process of the FHMM, assuming that the addition of the p_g term does not affect the best path chosen through the lattice, and therefore the value of u_t above. But, as the p_g term does not depend upon the value of u_t , every path in the lattice should be affected equally, and therefore the best path should remain the same.

However, there are other differences of implementation between the global subordinate-GMM FHMM and the output fusion presented above that make them slightly different for the purposes of these experiments. For the two products in (16) above to be combined to form (17), they must be multiplying over the same range of t -values, which is not the case here due to the different frame rates of each modality. Additionally, the normalisation performed in the FHMM nodes and also in the output fusion occur at different levels, introducing differences. Nevertheless, these factors could be easily controlled for, allowing output fusion to work as well as the global-subordinate-GMM-based FHMM model.

In a similar manner to this, the word and word-state subordinate-GMM-based FHMM models could be viewed as almost equivalent to HMM-GMM output fusion, provided that the sequence is first segmented into words or word-states, respectively, using the underlying HMM, and the correct subordinate GMM is

chosen for each segment. This is effectively what the FHMM model is doing with the significant difference being that the score-fusion occurs within the Viterbi process, so that the boundaries of the words or word-states have the possibility of moving based upon the subordinate observations. It is not clear at this stage how much this is in effect, and this will be covered in a future paper in more detail.

5.4 Results

DET plots showing the performance of our audio- and video-biased FHMM structures are shown in Figure 4. By comparing to the output fusion of the audio and video HMM, shown in both plots, it can be seen that the audio-biased structure is clearly more powerful than the video-biased version.

For the video-biased FHMMs, the word and word-state subordinate models fare considerably worse than the global subordinate model. As the global-subordinate-model can be replicated with output fusion, as discussed in the previous section, there is therefore little need of video-biased FHMMs in this situation. However, for audio-biased FHMMs there does appear to be a small benefit in using the word-state, or word FHMM over the global FHMM, particularly around the equal-error-rate region.

The main reason for the difference in performance between the two FHMM configurations is the ability of the dominant HMM to reliably estimate its underlying state sequence. The performance of the audio-biased FHMM shows that the audio HMM can reliably segment the sequences into sections of similar video appearance, but the video HMM does not appear able to locate segments of similar audio activity. Although the performance increase in this case is not large, the improved performance of the word-state FHMM over the global FHMM does appear to show that it is taking advantage of a temporal relationship between the audio states and video features.

6 Conclusion and Future Research

In this paper we have examined output fusion using both HMM and GMM classifiers in both the audio and video modalities and found that although temporal video information is clearly useful for lip-based speaker recognition using video HMMs, under output fusion most of this information appears to be lost. The performance of output fusion appears to be based mostly on the audio-classifier chosen, with the HMM performing better, and the choice of video classifier appears to only have a minor effect.

In an attempt to take greater advantage of the temporal video information in fusion with the audio, we adapted Pan et al.'s (2004) FHMMs to improve the robustness of the subordinate models to within-speaker variations, particularly on data recorded over multiple sessions. We found that our continuous FHMM model took advantage of the temporal relationship between the video observations and audio states to improve performance over the best performing output fusion in an audio-biased configuration. However, we found that the video-biased configuration showed no useful relationship between audio observations and video states.

In the audio-biased FHMM structure, a large portion of the video subordinate-GMMs are used to recognise primarily static features, such as lip or skin colour, which do not change throughout the sequence. As this type of information cannot form a temporal relationship with audio states, its effect on the subordinate-GMMs may be swamping the more dynamic information available in the movement of the

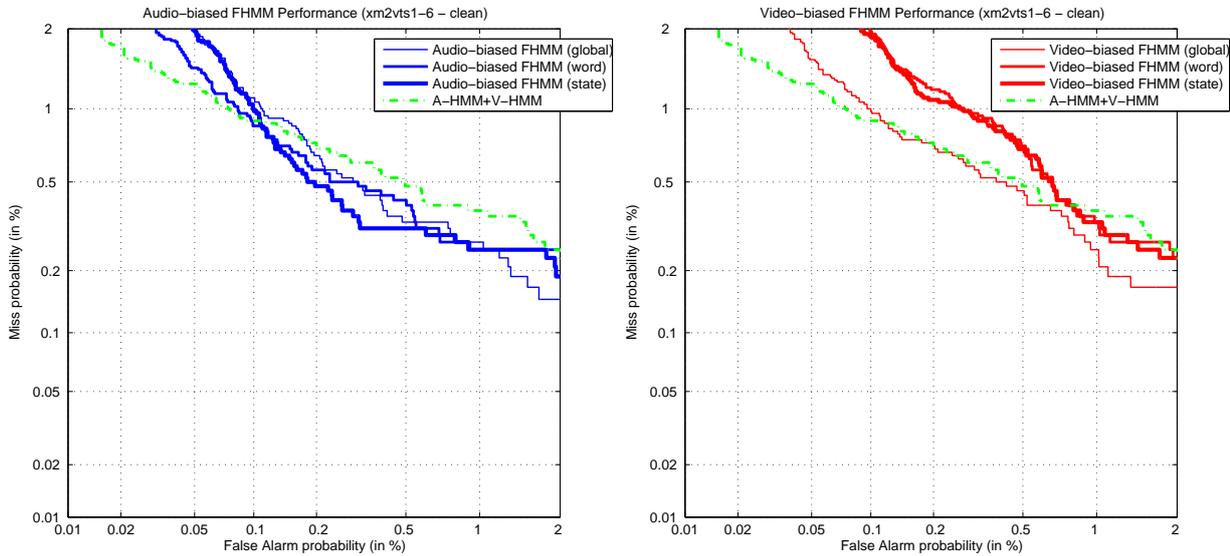


Figure 4: Detection error trade-off (DET) plots for FHMM speaker verification. (Note that the scale has changed from Figure 3.)

lips that could provide an improvement in the FHMM structure. A more efficient FHMM structure may be able to be realised by using more dynamic video features, and then performing output-fusion with a simple classifier using the static features so that the static information is not lost completely. Methods such as mean-image removal, optical flow or contour-based lip representations should provide better features to model the dynamic nature of visual speech.

Additionally, FHMMs should prove quite useful in other areas relating to audio-visual speech, such as speech recognition, or speaker detection.

7 Acknowledgments

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in (Messer et al. 1999) or at <http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/>.

References

- Brand, M. (1999), A bayesian computer vision system for modeling human interactions, in 'ICVS'99', Gran Canaria, Spain.
- Chibelushi, C., Deravi, F. & Mason, J. (2002), 'A review of speech-based bimodal recognition', *Multimedia, IEEE Transactions on* **4**(1), 23–37.
- Dean, D., Wark, T. & Sridharan, S. (2006), An examination of audio-visual fused HMMs for speaker recognition, in 'MMUA 2006', Toulouse, France.
- Heckmann, M., Kroschel, K., Savariaux, C. & Berthommier, F. (2002), DCT-based video features for audio-visual speech recognition, in 'International Conf. on Spoken Language Processing', Denver, Colorado, pp. 92093–0961.
- Lee, C.-H. & Gauvain, J.-L. (1993), Speaker adaptation based on MAP estimation of HMM parameters, in 'Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on', Vol. 2, pp. 558–561 vol.2.
- Luetttin, J. & Maitre, G. (1998), Evaluation protocol for the extended M2VTS database (XM2VTSDB), Technical report, IDIAP.
- Messer, K., Matas, J., Kittler, J., Luetttin, J. & Maitre, G. (1999), XM2VTSDB: The extended M2VTS database, in 'Audio and Video-based Biometric Person Authentication (AVBPA '99), Second International Conference on', Washington D.C., pp. 72–77.
- Nefian, A. V., Liang, L. H., Fu, T. & Liu, X. X. (2003), A bayesian approach to audio-visual speaker identification, in 'Audio-and Video-Based Biometric Person Authentication (AVBPA 2003), 4th International Conference on', Vol. 2688 of *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, Guildford, UK, pp. 761–769.
- Pan, H., Levinson, S., Huang, T. & Liang, Z.-P. (2004), 'A fused hidden markov model with application to bimodal speech processing', *IEEE Transactions on Signal Processing* **52**(3), 573–581.
- Pan, H., Liang, Z.-P. & Huang, T. S. (2001), 'Estimation of the joint probability of multisensory signals', *Pattern Recognition Letters* **22**(13), 1431–1437.
- Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2002), *The HTK Book*, 3.2 edn, Cambridge University Engineering Department., Cambridge, UK.

Voiceless Speech Recognition Using Dynamic Visual Speech Features

Wai Chee Yau Dinesh Kant Kumar Sridhar Poosapadi Arjunan

School of Electrical and Computer Engineering, RMIT University
GPO Box 2476V Melbourne, Victoria 3001, Australia
Email: waichee@ieee.org

Abstract

This paper describes a voiceless speech recognition technique that utilizes dynamic visual features to represent the facial movements during phonation. The dynamic features extracted from the mouth video are used to classify utterances without using the acoustic data. The audio signals of consonants are more confusing than vowels and the facial movements involved in pronunciation of consonants are more discernible. Thus, this paper focuses on identifying consonants using visual information. This paper adopts a visual speech model that categorizes utterances into sequences of smallest visually distinguishable units known as visemes. The viseme model used is based on the viseme model of Moving Picture Experts Group 4 (MPEG-4) standard. The facial movements are segmented from the video data using motion history images (MHI). MHI is a spatio-temporal template (grayscale image) generated from the video data using accumulative image subtraction technique. The proposed approach combines discrete stationary wavelet transform (SWT) and Zernike moments to extract rotation invariant features from the MHI. A feedforward multilayer perceptron (MLP) neural network is used to classify the features based on the patterns of visible facial movements. The preliminary experimental results indicate that the proposed technique is suitable for recognition of English consonants.

Keywords: visual speech recognition, wavelet transform, feedforward neural network.

1 Introduction

Speech recognition has been an important research subject that spans across multiple disciplines such as human-computer interaction (HCI), signal processing, linguistic and machine learning. Enormous research efforts are put into developing intelligent machines that are capable of comprehending utterances. Such speech-based devices are useful as they provide the flexibility for users to control computers using human speech.

However, the performance of the current speech recognition systems are still far behind as compare to human's cognitive ability in perceiving and understanding speech (Lippmann 1997). The main difficulty of the conventional speech recognition techniques based on audio signals is that such systems are sensitive to signal strength, ambient noise

and acoustic conditions. To overcome this limitation, the non acoustic speech modalities can be used to complement the audio signals. There are a number of options available such as visual (Goecke & Millar 2003, Potamianos, Neti, Huang, Connell, Chu, Libal, Marcheret, Haas & Jiang 2004), recording of vocal cords movements through electroglottograph (EGG) (Dikshit & R.W.Schubert 1995), mechanical sensing of facial movement and movement of palate, recording of facial muscle activity (Arjunan, Kumar, Yau & Weghorn 2006), facial plethysmogram and measuring the intra-oral pressure (Soquet, Saerens & Lecuit 1999). Vision-based speech recognition techniques are least intrusive and non invasive and this paper reports on such a technique for HCI application.

In our normal communication, the visual modality of speech is often incorporated into audio speech recognition (ASR) systems because the visual speech signals are invariant to acoustic noise and style of speech. Such systems that combine the audio and visual modalities to identify utterances are known as audio-visual speech recognition (AVSR) systems. AVSR systems can enhance the performance of the conventional ASR system especially under noisy condition (Chen 2001). Research where these AVSR systems are being made more robust, and able to recognize complex speech patterns of multiple speakers are being reported (Potamianos et al. 2004, Liang, Liu, Zhao, Pi & Nefian 2002). While AVSR systems are useful for applications such as for telephony in noisy environment, these are not suitable for people with speech impairment that have difficulty in producing speech sounds. AVSR systems are also not useful in situations where it is essential to maintain silence. Thus, the need for a voiceless, visual-only communication system arises. Such a system is also commonly known as lipreading or visual speech recognition or speechreading system.

Speechreading systems use the visual information extracted from the image sequence of the mouth to identify utterances. The visual speech information refers to the movement of the speech articulators such as the lips, facial muscles, tongue, teeth and jaw of the speaker. The complex range of reproducible sounds produced by people is a clear demonstration of the dexterity of the human mouth and lips- the key speech articulators. The possible advantages of such voiceless systems are (i) not sensitive to audio noise and change in acoustic conditions (ii) does not require the user to make a sound and (iii) suitable for users with speech impairment.

The visual cues contain far less classification power for speech compared to audio data and hence it is to be expected that speechreading systems would have only a small vocabulary. Such systems are also known to be user dependent, and hence it is important for such a system to be easy to train for a new user. And

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

because there is no audio cue, it is highly desirable that the system provide the user with active feedback to avoid any error in communication.

The main limitation with current speechreading systems is that these systems adopt a 'one size fits all' approach. Due to the large variation in the way people speak English, especially if we transgress the national and cultural boundaries, these have very high error rate, with error of the order of 90% for large vocabulary systems (Potamianos, Neti, Gravier & Senior 2003, Hazen 2006) and error rates in the range of 55% to 90% for small vocabulary system (Matthews, Cootes, Cox, Harvey & Bangham 1998), which demonstrates the inability of these systems to be used as voiceless, speech-controlled human computer interfaces.

What is required is a speaker-dependent system that is easy to train for individual users, with low computational complexity and can provide active feedback to the user. The system needs to be robust under changing conditions such as angle and distance of the camera, and insensitive to factors such as different skin color, texture and rapidity of speech of the speaker.

To achieve the above mentioned goals, this paper proposes a system where the camera is attached in place of the microphone to the commonly available head-sets. The advantage of this is that using this, it is no longer required to identify the region of interest, reducing the computation required. The video processing proposed is the use of accumulative image subtraction technique to directly segment the facial movements of the speaker. The proposed technique uses dynamic visual speech features based on the movements of the lower face region such as movements of the mouth, jaw and facial muscles. The proposed motion segmentation approach is based on the use of motion history images (MHI) where the video data is multiply with a ramp function and temporally integrated with greater weight to the recent movements. The resultant MHI is a 2-D grayscale image which is suitable for representing short duration complex movements of the lower face. Section 2 discusses on related work in the field of speechreading and Section 3 describes our proposed approach. Section 4 presents the methodology and Section 5 reports on the results of the initial experiments. Section 6 presents the discussion and findings based on the experimental results and Section 7 discusses the recommendations for possible future work.

2 Related Work

Numerous speechreading techniques are reported in the literature and comprehensive reviews on speech recognition research can be found in (Potamianos et al. 2003, Chen 2001, Stork & Hennecke 1996). Visual features used in speechreading systems can be divided into two main categories - shape-based and intensity-based. The shape-based features rely on the geometric shape of the mouth and lips and can be represented by a small number of parameters. The first speechreading system was proposed by Petajan (Petajan 1984) using shape-based features such as height, width and area of the mouth derived from the binary mouth images. Shape based features based on 3D coordinates of feature points (lip corners and midpoints of upper and lower lip) are extracted from stereo images in (Goecke & Millar 2003). Lip contours extracted using deformable template techniques such as active shape models (ASM) are used as visual speech features in (Matthews et al. 1998, Perez, Frangi, Solano & Lukas 2005). ASM obtains the lip information by fitting a statistical shape model of the

lip to the video frames. While such top-down, model-based approaches are less sensitive to the view angle of the camera and image noise, they rely only on the shape of the lip contours and do not contain information of other speech articulators. An extension to the ASM technique is active appearance model (AAM) that combines the shape model with a statistical model of the grey levels in the mouth region. The performance of AAM is demonstrated to outperform ASM in lip tracking (Matthews et al. 1998). However, both AAM and ASM techniques are sensitive to tracking error and modelling error.

Intensity-based features are derived directly from the pixel intensity values of the image around the mouth area (Liang et al. 2002, Potamianos et al. 2004, Hazen 2006, Saenko, Darrell & Glass 2004). Such features are extracted using bottom-up approach. The advantage of intensity-based systems is that accurate tracking and modelling of the lips are not required as opposed to model-based systems. The training of the statistical model of the lips is also not necessary for intensity-based approach thereby reducing the computational complexity of the systems. Intensity-based features are capable of representing visual information within the mouth cavity and also surrounding face region that are not represented in the high-level, shape-based features and lip contours (Potamianos et al. 2004). Nonetheless, the intensity-based features have much higher dimensionality if taking directly all the pixels from the mouth images. Dimensionality reduction or feature extraction techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) can be applied on the images to reduce the dimension of such features. The intensity-based features are demonstrated to yield better performance than shape-based features extracted using ASM and AAM algorithms in (Matthews et al. 1998). Similarly, intensity-based features using Discrete Cosine Transform (DCT) is also shown to outperform model-based features obtained using ASM algorithm in (Perez et al. 2005). This paper reports on the use of intensity-based features extracted from the MHI to represent facial movements for consonants recognition.

3 Theory

3.1 Visual Speech Model - Viseme Model

Human speech is organized as sequences of basic unit of speech sounds known as phoneme. Phonemes can be further dichotomized into vowels and consonants. The audio signals of consonants are less distinguishable than vowels (Chen 2001). Hence, the visual speech information is crucial in differentiating the consonants, especially in conditions where the acoustic signal strength is low or contaminated by noise. This paper focuses on the recognition of consonants due to the fact that consonants are easier to "see" and harder to "hear" than vowels (Kaplan, Bally & Garretson 1999). The pronunciation of vowels are produced with an open vocal tract whereas the production of consonants involve constrictions at certain part of the vocal tract by the speech articulators. Hence, the facial movements involved in pronunciation of consonants are more discernible than vowels. To represent the different facial movements when uttering consonants, a visual speech model is required.

This paper uses visemes to model visual speech. The motivation of using viseme as the recognition unit is because visemes can be concatenated to form words and sentences, thus providing the flexibility for the proposed visual speech recognition system to be extended into a large vocabulary system. The to-

tal number of visemes is much less than phonemes because speech is only partially visible (Hazen 2006). While the video of the speaker’s face shows the movement of the lips and jaw, the movements of other articulators such as tongue and vocal cords are often not visible. Hence, each viseme can correspond to more than one phoneme, resulting in a many-to-one mapping of phonemes-to-visemes.

Various viseme models had been proposed for AVSR applications (Hazen, Saenko, La & Glass 2004, Potamianos et al. 2004, Gordan, Kotropoulos & Pitas 2002). There is no definite consensus about how the sets of visemes in English is constituted (Hazen 2006). The number of visemes for English varies depending on factors such as the geographical location, culture, education background and age of the speaker. The geographic differences in English is most obvious where the sets of phonemes and visemes changes for different countries and even for areas within the same country. It is difficult to determine an optimal and universal viseme set that is suitable for all users. This paper adopts the viseme model established for facial animation applications by an international audiovisual object-based video representation standard known as MPEG-4. The motivation of using this model is because this enable the proposed system to be coupled with any MPEG-4 supported facial animation systems to form an interactive speech recognition and synthesis human computer interface. Based on the MPEG-4 viseme model, there is nine visemes associated with all English consonants. This paper adopts this nine visemes to represent the different facial movements when pronouncing consonants. The consonants chosen for experiments for each of the nine visemes are highlighted in bold fonts in Table 1.

Viseme Number	Phonemes	Example words
1	/p/,/b/,/m/	put, bed, me
2	/f/,/v/	far, voice
3	/th/,/dh/	think, that
4	/t/,/d/	tick, door
5	/k/,/g/	kick, gate
6	/sh/, /j/, /ch/	she, join, chair
7	/s/,/z/	sick, zeal
8	/n/,/l/	new, less
9	/r/	rest

Table 1: Viseme model of the MPEG-4 standard for English consonants.

3.2 Segmentation of the Facial Movements

In the proposed approach, the dynamic visual speech features which comprise of the facial movements of the speaker are segmented from the video data using a view-based approach named motion history images (MHI) (Bobick & Davis 2001). MHI is a spatial-temporal template that shows where and when movement of speech articulators (lips, teeth, jaw, facial muscles and tongue) occurs in the image sequence. MHI is generated using difference of frames (DOF) from the video of the speaker. Accumulative image subtraction is applied on the image sequence by subtracting the intensity values between successive frames to generate the difference of frames (DOFs). The delimiters for the start and stop of the motion are manually inserted into the image sequence of every articulation. The MHI of the video of the lips would have pixels corresponding to the more recent mouth movement brighter with larger intensity values. The intensity value of the MHI at pixel location

(x, y) of time t (or the t^{th} frame) is defined by

$$MHI_t = \max \bigcup_{t=1}^{N-1} B(x, y, t) \times t \quad (1)$$

N is the total number of frames used to capture the mouth motion. $B(x, y, t)$ is the binarisation of the DOF using the threshold a and $B(x, y, t)$ is given by

$$B(x, y, t) = \begin{cases} 1 & \text{if } Diff(x, y, t) \geq a, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

a is the predetermined threshold for binarisation of the DOF represented as $Diff(x, y, t)$. The value for the fixed threshold, a is optimized through experimentation. The DOF of the t^{th} frame is defined as

$$Diff(x, y, t) = |I(x, y, t) - I(x, y, t - 1)| \quad (3)$$

$I(x, y, t)$ represents the intensity value of pixel location with coordinate (x, y) at the t^{th} frame of the image sequence. In Eq. (1), the binarised version of the DOF is multiplied with a linear ramp of time to implicitly encode the timing information of the motion into the MHI (Kumar & Kumar 2005). By computing the MHI values for all the pixels coordinates (x, y) of the image sequence using Eq. (1) will produce a scalar-valued grayscale image (MHI) where the brightness of the pixels indicates the recency of motion in the image sequence. The proposed motion segmentation approach is computationally simple and is suitable for real time implementation. Figure 1 shows examples of 3 MHIs generated from the video of the speaker and Figure 2 illustrates the 9 MHIs that form the viseme model of MPEG-4 for English consonants used in the experiments.

The motivation of using MHI in visual speech recognition is the ability of MHI to remove static elements from the sequence of images and preserve the short duration facial movements. MHI is also invariant to the skin color of the speakers due to the DOF and image subtraction process involved in the generation of MHI.

3.2.1 Variation in Speed of Speech

The speed of phonation of the speaker might varies for each pronunciation of a phone. Hence, the speed of the mouth movements when the speaker is pronouncing a consonant might be different for each video recording. The variation in the speed of utterance results in the variation of the overall duration and there maybe variation in the microphases of the utterances. The details of such variations are difficult to model due to the large inter-subject and inter-experiment variations. This paper suggests a model to approximate such variations by normalizing the overall duration of the utterance. This is achieved by normalizing the intensity values of the MHI to in between 0 and 1 to minimize the difference in MHIs produced from video data of different rapidity of speech.

3.2.2 Issues Related to the Segmentation of the Facial Movements

MHI is a view sensitive motion representation technique. Therefore the MHI generated from the sequence of images is dependent on factors such as:

1. position of the speaker’s mouth normal to the camera optical axis
2. orientation of the speaker’s face with respect to the video camera

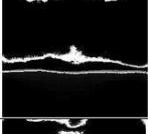
Consonants	Start Frame	Middle Frame	End Frame	MHI
/v/				
/m/				
/g/				

Figure 1: Examples of MHI generated from video of a speaker uttering three different consonants.

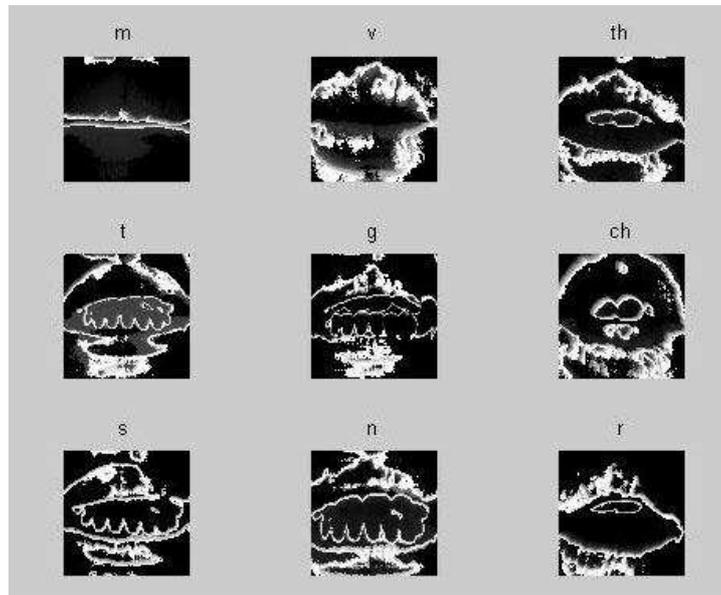


Figure 2: MHI of the 9 consonants from the MPEG-4 viseme model

- distance of the speaker's mouth from the camera (which changes the scale/size of the mouth in the video data)
- small variation of the mouth movement of the speaker while uttering the same consonant

This paper proposes the use of approximate image of discrete stationary wavelet transform (SWT) to obtain a time-frequency representation of the MHI that is insensitive to small variations of the mouth and lip movement. The proposed technique adopts Zernike moments as the region-based features to represent the SWT approximate image of the MHI to further reduce the dimension of the data. Zernike moments are chosen because they can be normalized to achieve rotation invariance.

3.2.3 Discrete Stationary Wavelet Transform

This paper proposes the use of discrete stationary wavelet transform (SWT) to obtain a transform representation of the MHI that is insensitive to small variations of the mouth and lip movement. While the classical discrete wavelet transform (DWT) is suitable for this, DWT results in translation variance (Mallat

1998) where a small shift of the image in the space domain will yield very different wavelet coefficients. The translation sensitivity of DWT is caused by the aliasing effect that occurs due to the downsampling of the image along rows and columns (Simoncelli, Freeman, Adelson & Heeger 1992). SWT restores the translation invariance of the signal by omitting the downsampling process of DWT, and results in redundancies.

2-D SWT at level 1 is applied on the MHI to produce a spatial-frequency representation of the MHI. The 2-D SWT is implemented by applying 1-D SWT along the rows of the image followed by 1-D SWT along the columns of the image. SWT decomposition of the MHI generates four images, namely approximation (LL), horizontal detail coefficients (LH), vertical detail coefficients (HL) and diagonal detail coefficients (HH) through iterative filtering using low pass and high pass filters. The approximate image is the smoothed version of the MHI and carries the highest amount of information content among the four images. LH, HL and HH sub images show the fluctuations of the pixel intensity values in the horizontal, vertical and diagonal directions respectively. The image moments features are computed from the ap-

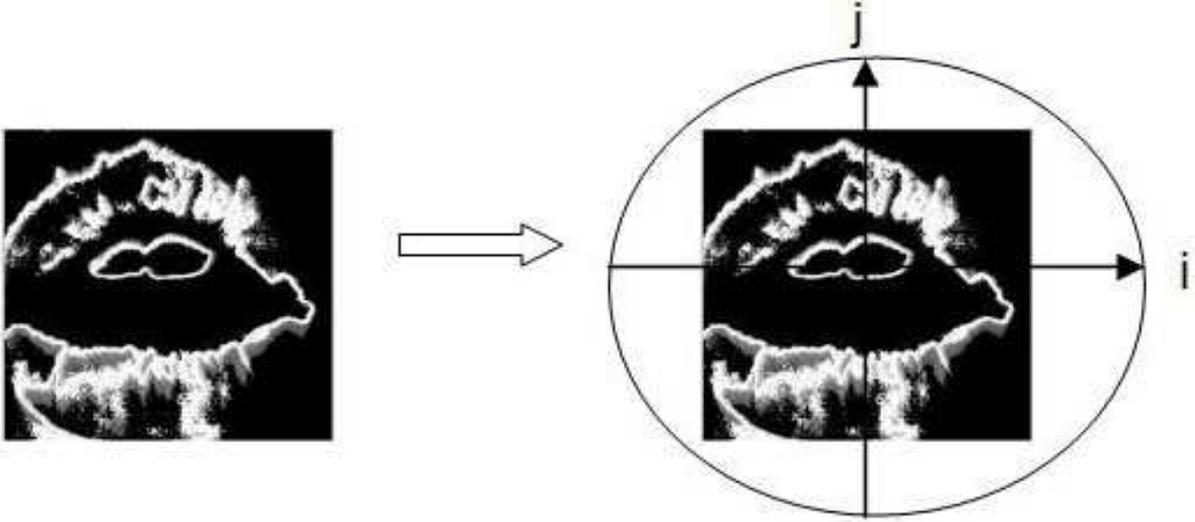


Figure 3: The square-to-circular transformation of the SWT approximation of MHI

proximate sub image.

Haar wavelet has been selected due to its spatial compactness and localization property. Another advantage is the low mathematical complexity of this wavelet. Compact features have to be extracted from the approximation (LL) to further reduce the size of the data. Since the gray levels of MHI are the temporal descriptors of motion occurring in the image sequence, thus it is intuitive to use global region-based feature descriptors to represent the approximation of the MHI. The proposed technique adopts Zernike moments as the region-based features to represent the SWT approximate image of the MHI.

3.3 Visual Speech Features - Zernike Moments

Zernike moments are image moments commonly used in recognition of image patterns (Khontazad & Hong 1990, Teague 1980). Zernike moments have been demonstrated to outperformed other image moments such as geometric moments, Legendre moments and complex moments in terms of sensitivity to image noise, information redundancy and capability for image representation (Teh & Chin 1988). The proposed technique uses Zernike moments as visual speech features to represent the SWT approximate image of the MHI.

Zernike moments are computed by projecting the image function $f(x, y)$ onto the orthogonal Zernike polynomial V_{nl} of order n with repetition l is defined within a unit circle (i.e.: $x^2 + y^2 \leq 1$) as follows:

$$V_{nl}(\rho, \theta) = R_{nl}(\rho)e^{-j\hat{l}\theta}; \hat{j} = \sqrt{-1} \quad (4)$$

where R_{nl} is the real-valued radial polynomial

The main advantage of Zernike moments is the simple rotational property of the features (Khontazad & Hong 1990). Zernike moments are also independent features due to the orthogonality of the Zernike polynomial V_{nl} (Teh & Chin 1988). $|l| \leq n$ and $(n - |l|)$ is even. Zernike moments Z_{nl} of order n and repetition l is given by

$$Z_{nl} = \left[\frac{n+1}{\pi} \right] \int_0^{2\pi} \int_0^{\infty} [V_{nl}(\rho, \theta)] f^*(\rho, \theta) \rho d\rho d\theta \quad (5)$$

$f(\rho, \theta)$ is the intensity distribution of the approximate image of MHI mapped to a unit circle of radius ρ and angle θ where $x = \rho \cos\theta$ and $y = \rho \sin\theta$.

For the Zernike moments to be orthogonal, the approximate image of the MHI is scaled to be within a unit circle centered at the origin. The unit circle is bounded by the square approximate image of the MHI. The center of the image is taken as the origin and the pixel coordinates are mapped to the range of the unit circle i.e.: $x^2 + y^2 \leq 1$. Figure 3 shows the square-to-circular transformation performed for the computation of the Zernike moments that transform the square image function ($f(x, y)$) in terms of the x-y axes to a circular image function ($f(\rho, \theta)$) in terms of the i-j axes.

To illustrate the rotational characteristics of Zernike moments, consider β as the angle of rotation of the image. The resulting rotated Zernike moment Z'_{nl} is

$$Z'_{nl} = Z_{nl}e^{-il\beta} \quad (6)$$

Z_{nl} is the Zernike moment of the original image. Eq. (6) demonstrates that rotation of an image results in a phase shift on the Zernike moments (Teague 1980). The absolute value of Zernike moments are rotation invariant (Khontazad & Hong 1990) as shown in the equation below

$$|Z'_{nl}| = |Z_{nl}| \quad (7)$$

This paper uses the absolute value of the Zernike moments, $|Z'_{nl}|$ as the rotation invariant features of the SWT of MHI. By including higher order moments, more information of the MHI can be represented by the Zernike moments features. However, this inherently increases the size of the features and makes it prone to noise. An optimum number of Zernike moments need to be selected to trade-off between the dimensionality of the feature vectors and the amount of information represented by the features. 49 Zernike moments that comprise of 0th order moments up to 12th order moments have been used as features to represent the approximate image of the MHI for each consonant. Table 2 lists the 49 Zernike moments used in the experiments.

Order	Moments	No. of Moments
0	Z_{00}	1
1	Z_{11}	1
2	$Z_{20} Z_{22}$	2
3	$Z_{31} Z_{33}$	2
4	$Z_{40} Z_{42} Z_{44}$	3
5	$Z_{51} Z_{53} Z_{55}$	3
6	$Z_{60} Z_{62} Z_{64} Z_{66}$	4
7	$Z_{71} Z_{73} Z_{75} Z_{77}$	4
8	$Z_{80} Z_{82} Z_{84} Z_{86} Z_{88}$	5
9	$Z_{91} Z_{93} Z_{95} Z_{97} Z_{99}$	5
10	$Z_{10,0} Z_{10,2} Z_{10,4} Z_{10,6} Z_{10,8} Z_{10,10}$	6
11	$Z_{11,1} Z_{11,3} Z_{11,5} Z_{11,7} Z_{11,9} Z_{11,11}$	6
12	$Z_{12,0} Z_{12,2} Z_{12,4} Z_{12,6} Z_{12,8} Z_{12,10} Z_{12,12}$	7

Table 2: List of the 49 Zernike Moments and Their Corresponding Number of Features From Order Zero to Order Twelve

3.4 Classification Using Feedforward Neural Network

There are a number of possible classifiers that maybe suitable for such a system. The selection of the appropriate classifier would require statistical analysis of the data that would also identify the features that are irrelevant. Supervised neural network approach lends itself for identifying the separability of data even when the statistical properties and the types of separability (linear or nonlinear) is not known and without even requiring the estimating of the kernel. While it may be suboptimum, it is an easy tool to implement as a first step.

This paper presents the use of artificial neural network (ANN) to classify Zernike moments features into one of the class of consonants. ANN has been selected because it can solve complicated problems where the description for the data is not easy to compute. The other advantage of the use of ANN is its fault tolerance and high computation rate due to the massive parallelism of its structure (Kulkarni 1994). The functionality of the ANN to be less dependent on the underlying distribution of the classes as opposed to other classifiers such as Bayesian classifier and Hidden Markov Models (HMM) is yet another advantage for using ANN in this application (Stork & Hennecke 1996).

A supervised feed-forward multilayer perceptron (MLP) ANN classifier with back propagation (BP) learning algorithm is integrated in the visual speech recognition system described in this paper. The ANN is provided with a number of training vectors for each class during the training phase. MLP ANN was selected due to its ability to work with complex data compared with a single layer network. Due to the multilayer construction, such a network can be used to approximate any continuous functional mapping (Bishop 1995). This paper proposes the use of a three-layer network with BP learning algorithm to classify the visual speech features. The advantage of using BP learning algorithm is that the inputs are augmented with hidden context units to give feedback to the hidden layer and extract features of the data from the training events (Haug 2001). Trained ANNs have very fast classification speed (Freeman & Skapura 1991) thus making them an appropriate classifier choice for real time visual speech recognition applications. Figure 4 shows the overall block diagram of the proposed technique.

4 Experiments

Experiments were conducted to evaluate the performance of the proposed visual speech recognition techniques in classifying English consonants. The experiments were approved by the Human Experiments Ethics Committee of the university. Nine consonants that form the viseme model of English consonants according to the MPEG-4 standard are tested in the experiment. The nine consonants tested (/m/, /v/, /th/, /t/, /g/, /ch/, /s/, /n/ and /r/) were highlighted in bold in Table 1.

4.1 Video Recording and Processing

Video data was recorded from one speaker using an inexpensive web camera in a typical office environment. This was done towards having an inexpensive and practical voiceless communication system using low resolution video recordings. The video camera focused on the mouth region of the speaker and the camera was kept stationary throughout the experiment. The following factors were kept the same during the recording of the videos: window size and view angle of the camera, background and illumination. 20 video data of size 240 x 240 was recorded for each of the nine consonants. Thus, a total of 180 video data was created. The video data was stored as true color (.AVI) files and every AVI file had a duration of two seconds to ensure that the speaker had sufficient time to utter each of the consonant. The frame rate of the AVI files was 30 frames per second. One MHI was generated from each of the AVI file. An example of MHI for each of the nine consonants are shown in Figure 3.

4.2 Features Extraction

SWT at level-1 using Haar wavelet was applied on the MHIs and the approximate image (LL) was used for analysis. Zernike moments are computed from circular region of interest while the MHI is a square image. Hence, square-to-circular transformation of the SWT approximate image of the MHI had been done to compute the orthogonal Zernike moments features. 49 Zernike moments that comprise of 0th order moments up to 12th order moments have been used as features to represent the SWT approximate image of the MHI for each consonants.

4.3 Classification

The next step of the experiments was to classify the features using artificial neural network (ANN), which

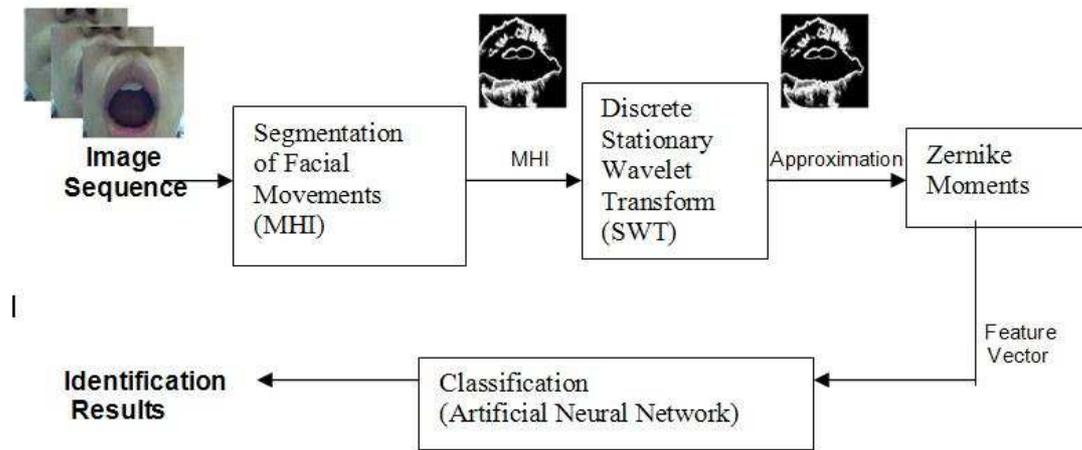


Figure 4: Block diagram of the proposed visual speech recognition approach.

Viseme	Recognition Rate
/m/	100%
/v/	87%
/th/	65%
/t/	74%
/g/	85%
/ch/	91%
/s/	93%
/n/	74%
/r/	93%

Table 3: Mean Classification Accuracies for 9 Visemes(Consonants).

can learn patterns of features with nonlinear separation. The Zernike moments features were fed to ANN to classify the features into one of the consonants. Multilayer perceptron (MLP) ANN with back-propagation (BP) learning algorithm was employed in the proposed system. The architecture of the ANN consisted of two hidden layers. The size of the input layer of the ANN was chosen to be same as the size of the features which was 49 nodes. The size of the output layer of the ANN was 9 which corresponded to the number of visemes (classes) available. The total numbers of hidden nodes was 140 which was determined iteratively through experimentation. Sigmoid function was the threshold function and the type of training algorithm for the ANN was gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima. In the experiments, Zernike moments features of 10 MHIs of each consonants were used to train the ANN. The remaining 10 MHIs (that were not used in training the ANN) were presented to the ANN to test the ability of the trained ANN in recognizing the nine consonants. The statistical mean and variance of the classification accuracies of the data are determined by repeating the experiments 10 times. For each repetition of the experiment, the 10 test samples for each consonants were selected randomly with different combinations (permutations) to train the ANN and the remaining 10 MHIs were used as test samples.

5 Results and Observations

The experiments have tested the robustness of the use of MHI features to identify the human speech visemes with a feedforward neural network as the classifier. The ANN used was trained for an individual subject. The mean recognition accuracies of the ANN for the 10 repetitions of the experiments are tabulated in Table 3. From this table, it is observed that the mean success rate for identifying the viseme based consonants is 84.7% with a standard deviation of 2.8%.

6 Discussion

The results indicate that the proposed technique based on dynamic visual speech information (facial movements) is suitable for consonants recognition. The results indicate that the different patterns of facial movements can be used to classify the 9 visemes of English consonants based on the MPEG-4 viseme model.

The good results demonstrate the ability of ANN to learn the patterns of the facial movement features. From the results, it is observed that a small number of samples are sufficient to suitably train the ANN based system, indicating the sufficient compactness of each class of the data.

One of the possible reason for the misclassifications of the test samples by the ANN can be attributed to the inability of vision-based technique to capture the occluded articulators movements. Example, the movement of the tongue within the mouth cavity is not visible (occluded by the teeth) in the video data during the pronunciation of /n/. Thus, the resultant MHI of /n/ does not contain information on the tongue movement.

While the error rates of the experiments are much lower than the 90% error reported by (Potamianos et al. 2003, Hazen 2006), the authors would like to point out that it is not appropriate to compare our results with other related work as this system has only been tested using a small vocabulary consisting of discrete phones of a single speaker. Other work has used a much larger vocabulary of continuous speech database of multiple speakers. Our system has been designed for specific applications such as control of machines using simple commands consisting of discrete utterances while other systems were developed for recognition of continuous speech. Nevertheless the

85% accuracies of our system is encouraging.

The authors suggest that one reason for the high accuracies of this system is that it is not only based on lip movement, but is based on the movement of the mouth, jaw and facial muscles. While lips are important articulators of speech, other parts of the mouth are also important, and this approach is closer to the accepted model of human visual speech perception.

The results demonstrate that a computationally inexpensive system which can easily be developed on a DSP chip can be used for such an application.

7 Conclusion

This paper reports on a voiceless speech recognition technique using video of the speaker's mouth that is computationally inexpensive and suitable for HCI applications. The proposed technique recognizes English consonants based on the dynamic speech information - facial movements of the speaker during phonation.

This paper adopts the MPEG-4 viseme model as the visual speech model to represent all the English consonants. An error rate of approximately 15% is obtained in classifying the consonants using this model. The misclassifications of the features can be attributed to the occlusion of speech articulators. Thus, non visible movements during the production of the consonants (such as movements of the tongue and vocal cords) are not represented in the visual speech features.

The results of our experiments suggest that the proposed technique is suitable in recognizing consonants using the information of the facial movements. The proposed system is easy to train for individual users and is designed for speaker-dependent speech-controlled applications. For future work, the authors intend to design a more suitable visual speech model for the consonants that accounts for the nonvisible articulators movements. Also, the authors intend to compare the performance of ANN with other classifiers such as Support Vector Machines (SVM) and Hidden Markov Models (HMM) to determine the optimum classifier for our application. Such a system could be used to drive computerized machinery in noisy environments. The system may also be used for helping disabled people to use a computer and for voice-less communication.

References

- Arjunan, S. P., Kumar, D. K., Yau, W. C. & Weghorn, H. (2006), Unspoken vowel recognition using facial electromyogram, *in* 'IEEE EMBC', New York.
- Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.
- Bobick, A. F. & Davis, J. W. (2001), 'The recognition of human movement using temporal templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 257–267.
- Chen, T. (2001), 'Audiovisual speech processing', *IEEE Signal Processing Magazine* **18**, 9–21.
- Dikshit, P. & R.W.Schubert (1995), Electroglossograph as an additional source of information in isolated word recognition, *in* 'Fourteenth Southern Biomedical Engineering Conference', LA, pp. 1–4.
- Freeman, A. & Skapura, M. (1991), *Neural Networks : Algorithms, Applications and Programming Techniques*, Addison-Wesley.
- Goecke, R. & Millar, J. B. (2003), Statistical analysis of the relationship between audio and video speech parameters for Australian English, *in* 'Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003', France, pp. 133–138.
- Gordan, M., Kotropoulos, C. & Pitas, I. (2002), Application of support vector machines classifiers to visual speech recognition, *in* 'International Conference on Image Processing', Vol. 3, Romania, pp. III–129 – III–132.
- Huang, K. Y. (2001), Neural network for robust recognition of seismic patterns, *in* 'IJCNN'01, Int Joint Conference on Neural Networks', Vol. 4, Washington, USA, pp. 2930–2935.
- Hazen, T. J. (2006), 'Visual model structures and synchrony constraints for audio-visual speech recognition', *IEEE Transactions on Audio, Speech and Language Processing* **14**(3), 1082–1089.
- Hazen, T. J., Saenko, K., La, C. H. & Glass, J. R. (2004), A segment-based audio visual speech recognizer : Data collection, development and initial experiments, *in* 'Int Conf on Multimodal Interfaces', State College, Pennsylvania, pp. 235–242.
- Kaplan, H., Bally, S. J. & Garretson, C. (1999), 'Speechreading: A way to improve understanding', pp. 14–16.
- Khontazad, A. & Hong, Y. H. (1990), 'Invariant image recognition by zernike moments', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 489–497.
- Kulkarni, A. D. (1994), *Artificial Neural Network for Image Understanding*, Van Nostrand Reinhold.
- Kumar, S. & Kumar, D. K. (2005), 'Visual hand gesture classification using wavelet transform and moment based features', *International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP)* **3**(1), 79–102.
- Liang, L., Liu, X., Zhao, Y., Pi, X. & Nefian, A. V. (2002), Speaker independent audio-visual continuous speech recognition, *in* 'IEEE Int. Conf. on Multimedia and Expo', Vol. 2, Switzerland, pp. 25–28.
- Lippmann, R. P. (1997), 'Speech recognition by machines and humans', *J. Speech Communication* **22**, 1–15.
- Mallat, S. (1998), *A Wavelet Tour of Signal Processing*, Academic Press.
- Matthews, I., Cootes, T., Cox, S., Harvey, R. & Bangham, J. A. (1998), Lipreading using shape, shading and scale, *in* 'Proc. Auditory-Visual Speech Processing', Terrigal, Australia, pp. 73–78.
- Perez, J. F. G., Frangi, F. A., Solano, E. L. & Lukas, K. (2005), Lip reading for robust speech recognition on embedded devices, *in* 'ICASSP'05, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing', Vol. 1, Philadelphia, PA, USA, pp. 473–476.
- Petajan, E. D. (1984), Automatic lip-reading to enhance speech recognition, *in* 'GLOBECOM'84, IEEE Global Telecommunication Conference'.

- Potamianos, G., Neti, C., Gravier, G. & Senior, A. W. (2003), Recent advances in automatic recognition of audio-visual speech, *in* 'Proc. of IEEE', Vol. 91, pp. 1306–1326.
- Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., Marcheret, E., Haas, N. & Jiang, J. (2004), Towards practical deployment of audio-visual speech recognition, *in* 'IEEE Int. Conf. on Acoustics, Speech, and Signal Processing', Vol. 3, Canada, pp. iii777–780.
- Saenko, K., Darrell, T. & Glass, J. (2004), Articulatory features for robust visual speech recognition, *in* 'ICMI'04', pp. 152–158.
- Simoncelli, E. P., Freeman, W. T., Adelson, E. H. & Heeger, D. J. (1992), 'Shiftable multiscale transform', *IEEE Transactions on Information Theory* **38**, 587–607.
- Soquet, A., Saerens, M. & Lecuit, V. (1999), Complementary cues for speech recognition, *in* '14th International Congress of Phonetic Sciences (ICPhs)', San Francisco, pp. 1645–1648.
- Stork, D. G. & Hennecke, M. E. (1996), Speechreading: An overview of image processing, feature extraction, sensory intergration and pattern recognition techniques, *in* '2nd International Conference on Automatic Face and Gesture Recognition (FG '96)', USA, pp. XVI–XXVI.
- Teague, M. R. (1980), 'Image analysis via the general theory of moments', *Journal of the Optical Society of America* **70**, 920–930.
- Teh, C. H. & Chin, R. T. (1988), 'On image analysis by the methods of moments', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 496–513.

ABSTRACTS

The papers in this section were submitted as extended abstracts that were not part of the peer review process.

Emotions in HCI – An Affective E-Learning System

Robin Kaiser, Karina Oertel

Fraunhofer Institute for Computer Graphics Rostock, Human Centered Interaction Technologies, Germany

mail: {robin.kaiser, karina.oertel}@igd-r.fraunhofer.de

Abstract

This paper presents results of a master thesis at the Fraunhofer Institute for Computer Graphics Rostock (IGD-R) at the department for Human-Centered Interaction. Using an emotion recognition sensor system, an e-learning system was enhanced with affective abilities. By taking certain actions, the user is supported to handle negative emotions, which should enable a better learning as well as a greater satisfaction. The affective communication and actions are encapsulated by an Affective Component, which was implemented as a prototype and evaluated at a first glance.

Keywords: HCI, Affective Computing, E-Learning

1 Introduction

Despite the ongoing development in technology over the past decades, computers still do not consider emotions of their users, even though many studies (e.g. Reeves & Nass 1996) showed how important they are for human-computer interaction. With the focus on innovative and user centred interaction technologies, the interplay between emotions and computers, widely known as affective computing (Picard 1997), plays an important role at the department for Human-Centered Interaction.

Traditional e-learning systems focus on the learning target only, whereas human expert tutors also concentrate on the emotional component of learning (Lepper & Chabay 1988). This seems to be a good model for e-learning, as negative emotions like boredom or anger reduce cognitive effort and in consequence hinder the achievement of learning goals.

2 Emotion detection

Emotion detection represents the first step in building affective applications. One way of detecting emotions is to analyse physiological data to deduce emotional states. The emotion recognition sensor system (EREC), developed at the IGD-R, consists of a sensor glove, a chest belt and a data collection unit (Figure 1).

Integrated in the glove are sensors for e.g. skin resistance and skin conductivity. Evaluated and enhanced sensor data are wirelessly transmitted and made available to a PC (Peter et. al. 2005). EREC is used for emotion detection by the Affective Component.



Figure 1: Emotion Recognition Sensor System (EREC)

3 Negative emotions and target emotions

The Affective Component is based on Russell's circumplex model of emotions (Russell 1980), a dimensional approach for classifying emotions. It assumes the existence of the dimensions valence and arousal utilized to describe different emotions. Instead of single emotions, only regions of the valence-arousal-space were taken into account. Thus, a concrete classification of emotions could be avoided.

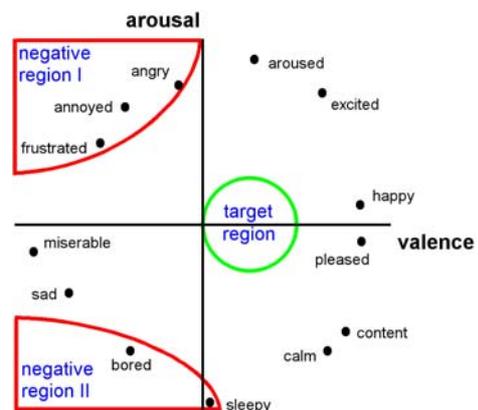


Figure 2: Russell's circumplex model with regions

For learning, two negative regions in the valence-arousal-space can be defined that should be avoided. By negative valence and positive arousal region I is described, which stands for emotions like frustration and anger. Emotions like boredom and sleepiness are represented by region II, located in an emotion-space

Copyright © 2006, Australian Computer Society, Inc. This paper appeared at the *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not for profit purposes permitted provided this text is included.

characterized by negative valence and negative arousal. The target emotional region, specified by a slight positive valence and neutral arousal, provides a maximum of efficiency and productivity in learning (Kaiser 2006). Besides, the user will feel more comfortable during the learning process.

4 Affective measures and procedure

A catalogue of affective measures describes actions to support the user in handling negative emotions. Besides a distinction of measures for both regions, measures are application-independent or application-dependent. Examples for application-independent measures used by the Affective Component are motivational statements, the possibility to express displeasure, the suggestion of short break or even a way to treat the computer with hammer, flamethrower and chain saw to reduce stress (Figure 3).

Application-dependent measures, bound to the given e-learning system or at least to the application domain, are a change of lesson, another way of presenting the subject (e.g. an animation instead of pure text) or the start of a questionnaire to check the learners learn progress.

With a technology for detecting different emotions and well-defined regions of negative and target emotions, it is still open what the affective procedure looks like. If negative emotions from one region are dominating for a certain time, an affective measure depending on the region is selected and suggested. If the user accepts, the chosen action is executed. Hopefully, his emotional state will be improved thereafter. For the pilot study, an intermediate step was needed. The correct detection of emotions was verified by asking the user, to ensure the initiation of a correct measure. However, it might be better to leave this out for final application.

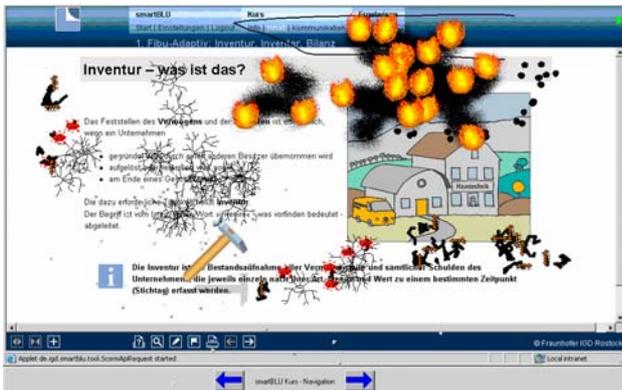


Figure 3: Measure to relieve stress and aggression (StressRelief 2006)

5 Pilot study and preliminary results

The implemented Affective Component was tested with a pilot study. SmartBLU (SmartBLU 2006), a learning management system, was used as underlying e-learning system. Three questions should be clarified. Are users more pleased when using the affective version of smartBLU? Is a greater success in learning achievable? Do users of the affective version of smartBLU stay less at region I and region II and stay longer at the target emotional region? The first two questions were proofed

by using questionnaires regarding satisfaction and factual knowledge respectively. Question three was checked by implementing the Affective Monitor, which logs the residence time at the different regions, the emotional states and the status of the Affective Component.

First findings show a tendency towards the expected results. Especially the success in learning of the affective testing group was slightly better than of the control group. However, results regarding the other questions were ambiguous. Possible causes may be measuring inaccuracies, the limited test duration and finally the limited number of test participants.

6 Conclusion and future work

The presented approach should only be considered as a first attempt of building an Affective Component making an e-learning system affective. Next steps are intended for further improvement. The selection of a certain measure, done coincidentally at the moment, could be based on information about the user. An affective user model is needed, which allows to arrange the single measures according to priority.

Furthermore, the affective procedure needs to be adapted individually. Based on experiences with the user, the moment of the initiation of measures, the minimum time between different measures or even the possibility to suggest an alternative measure in case of a rejection by the user could be defined more precisely. Finally, more extensive studies are needed for a final evaluation.

7 References

- Kaiser, R. (2006), Master thesis: Prototypische Entwicklung einer affektiven Komponente für ein E-Learning System (Prototypical development of an affective component for an e-learning system), University of Rostock, Germany, Department of Computer Science.
- Lepper, M.R. & Chabay R.W. (1988), Socializing the Intelligent Tutor: Bringing Empathy to Computer Tutors, in *Learning issues for intelligent tutoring systems*, Springer Verlag, New York, pp. 242-257.
- Peter, C. et al. (2005), A Wearable Multi-Sensor System for Mobile Acquisition of Emotion-Related Physiological Data, in *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction*, Beijing 2005. Springer Verlag Berlin, Heidelberg, New York, pp. 691-698.
- Picard, R.W. (1997), *Affective Computing*, MIT Press, Cambridge, Massachusetts.
- Reeves, B. & Nass, C. (1996), *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, New York.
- Russell, J.A. (1980), A circumplex model of affect, in *Journal of Personality and Social Psychology*, 39, pp. 1161-1178.
- SmartBLU (2006): Learning Management System smartBLU. <http://www.smartblu.de>. Accessed 30 Sep 2006.
- StressRelief (2006): Freeware, <http://www.gemtree.com/program.htm>. Accessed 30 Sep 2006.

Video to the Rescue

Girija Chetty and Michael Wagner

School of Information Sciences and Engineering, University of Canberra, Australia

Email: {girija.chetty, Michael.wagner}@canberra.edu.au

Abstract

Automatic person identity verification based on biometrics is a challenging problem, and has received much attention during recent years due to its many applications in on-line transaction processing, law enforcement, and security applications. However, most identity verification systems are primarily based on voice biometrics, and hence are more vulnerable to acoustic noise and channel distortions, in addition to train/test mismatch conditions. In this paper, we show how we can use video information to improve the performance of identity verification systems. The approach based on multimodal fusion of voice and face information from speaking face video allows robust identity verification performance. In addition, depending on the type of features and fusion technique used, it is also possible to perform liveness checks, allowing the system to detect fraudulent attacks on the system.

Keywords: identity verification, multimodal fusion, face-voice.

1 Introduction

Information about a person's identity is multimodal. Yet, most biometric based person identity verification systems limit themselves to only a single modality, such as person's voice. A good example of a system that combines multiple information sources is the human being, e.g. it has been shown that simultaneously seeing and listening a person talking greatly increases intelligibility. The audio-visual speech recognition approaches have exploited this fact and improved the performance of speech recognition systems (Potamianos et al. 2003, Chelubishi et al. 2003).

However, person identity verification systems have been mostly based on the voice modality, and though they achieve high performance when the audio signal-to-noise ratio (SNR) is high, the performance degrades quickly as the test SNR decreases or train/test mismatch increases (Ben Yacoub et al. 1999).

To combat the limitations of unimodal audio based identity verification approaches, multimodal approaches based on combining video information with acoustic

information, similar to speech recognition approaches can be used, and improvement in both robustness and overall performance can be achieved. In addition, based on the type of features extracted from several modes/sections of the video, such as the face, the voice and the mouth region, and the type of multimodal fusion technique used, it is also possible to verify 'liveness', which allows the system to thwart fraudulent attacks on the system, such as replay of client-specific information, including pre-recorded audio or video or still/ animated photo of the face.

The audio, face, and mouth modalities contain non-redundant, complementary information about person identity (Dieckmann et al. 1997). However, in order to exploit this, several issues need to be addressed, such as how to account for the reliabilities of the modalities, what type of features need to be extracted, and at what level to carry out the fusion. Only a few studies have investigated the combination of audio, face, and temporal mouth information for the purpose of person identity verification (Dieckmann et al. 1997, Yemez et al 2003). The issue of liveness verification has hardly been addressed in most identity verification studies, except some recent studies in (Chetty & Wagner 2004, Bredin & Cholet 2006). The majority of studies were unimodal using either the audio or static face modalities (Reynolds et al 1995, Rabiner 1989)

In this paper, we show how we can use video information from speaking faces to increase the identity verification performance. The results of two different types of experiments, the first type for identity verification, and the second for liveness verification, with three different speaking face corpora, VidTIMIT (Sanderson et al. 2005) AVOZES (Goecke 2004), and UCBN, shows a significant improvement in performance in terms of performance measures such as DET curves and EER rates, when video information was fused with audio information.



Figure 1: Faces from VidTIMIT, UCBN and AVOZES corpus

Figure 1, shows sample faces from VidTIMIT, UCBN and AVOZES corpus. The three types of corpora

Copyright © 2006, Australian Computer Society, Inc. This paper appeared at the *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

represent very different types of speaking face data, VidTIMIT with original audio recorded in a noisy environment and clean visual environment, UCBN with clean audio and visual environments, but complex visual backgrounds, and AVOZES with stereo face data for better 3D face modeling.

2 Multimodal fusion

Techniques for combining different information sources can be broadly grouped into feature fusion and late-fusion techniques. Late-fusion techniques combine information after mapping from the feature space to the opinion/decision space using either a classifier or an expert. With late fusion it is possible to combine opinions from different experts, even if their outputs are not commensurate (different range values). In feature fusion, information is combined before any use of classifier or expert. Feature fusion has been widely for instance in lip-reading where visual and speech features are combined to increase intelligibility. Feature fusion techniques are more appropriate when the information sources are closely synchronized, such as visual speech information from mouth region of a speaking face. By using late fusion for such tasks, many of the correlation properties of the joint audio-video data are lost. For these reasons, we have used features fusion to combine the features from voice and mouth region before the classification stage for liveness verification experiments.

3 Experiments

For identity verification experiments, for face and voice modality, we separately formulate the problem as a basic hypothesis test, where for voice mode, given a speech segment S , a decision whether it was spoken by person P_i has to be made. The optimum test is given by the log-likelihood ratio:

$$LLR(P_i) = \log \left\{ \frac{p(S/P_i)}{p(S/P_{BM})} \right\} \quad (1)$$

where $p(S/P_i)$ and $p(S/P_{BM})$ are the conditional probability density functions using the models of person P_i and background BM respectively, which are often modeled using Gaussian Mixture Models. The features used for modeling the speakers were MFCC acoustic and delta acoustic features(12+8) and PCA face features(20) after appropriate normalization and pre-processing. More details about features used and modeling of speakers can be found in (Chetty & Wagner 2004). After the speaker recognition process a confidence value $LLR(P_i)_a$ is available which can be fused with the face confidence value $LLR(P_i)_v$. In Figure 2, a scatter plot showing face and voice confidences in the likelihood space is shown. It can be seen that true and false candidates (impostors) are better classified in the two-dimensional space.

Our experiments show that the fusing video with audio information increases the average identity verification performance by about 50% compared to audio only approach to person recognition, shown in 1st row of Table 1.

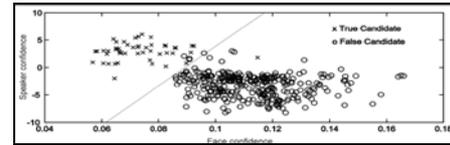


Figure 2: Scatter plot showing face and voice confidences in the likelihood space.

For liveness verification experiments, we built a lip-voice fusion vector by feature fusion of acoustic MFCC features (8) with PCA features from lip-region(3) and lip width/mouth height ratio(1) for building the speaker models. In the training phase, a 10-mixture Gaussian mixture model of each client's lip-voice fusion vector was built. In the test phase, clients' live test recordings were evaluated against a client's model λ by determining the log-likelihoods ($\log p(X|\lambda)$) of the time sequences X of lip-voice feature vectors. The protocol was extended for replay attack tests, by synthesizing a number of "fake" recordings by combining the sequence of audio feature vectors from each test utterance with ONE lip feature vector chosen from the sequence of lip feature vectors. The performance in terms of average EERs (Equal Error Rates) for three corpora achieved with separate audio, lip and lip-voice fusion vectors are shown in 2nd row of Table 1.

Experiments	Audio Only(EER)	Video Only(EER)	Audio+Video EER
Identity Verification	7.1%	5.54 %	3.6 %
Liveness Verification	6.86%	4.68%	3.39%

Figure 1: Identity and liveness verification performance

4 References

- Ben-Yacoub S., X Abdeljaoued V., and Mayoraz E. (1999), "Fusion of face and speech data for person identity verification," IEEE Transactions on Neural Networks, vol. 10, pp. 1065-1074.
- Bredin H., Chollet G.(2006), "Measuring Audio and Visual Speech Synchrony: Methods and Applications", Proc. International Conference on Visual Information Engineering VIE 2006, Bangalore, India,.
- Dieckmann U., Plankensteiner P., and Wagner T.(1997), "SESAM: A biometric person identification system using sensor fusion," Pattern Recognition Letters, vol. 18, pp. 827-833.
- Chetty G., and Wagner M.(2004), "Liveness Verification in Audio-Video Speaker Authentication", Proceedings of the 10th ASSTA conference.
- Goecke, R., and Millar J.B. (2004), "The Audio-Video Australian English Speech Data Corpus AVOZES", Proceedings of the 8th INTERSPEECH 2004 - ICSLP, Volume III, pages 2525-2528.
- Sanderson, C. and Paliwal K.K., "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 2409-2419, 2003.
- Yemez Y., Kanak A., Erzin E., and Tekalp A. M.(2003), "Multimodal speaker identification with audiovideo processing," presented at Proc. of the Int'l Conf. on Image Processing, Barcelona, Spain.

Author Index

Arjunan, Sridhar Poosapadi, 13, 93

Caelli, Terry, iii
Carreira, João, 73
Casas, Josep R., 19
Chetty, Girija, 107
Chik, Desmond, 61

Dean, David, 87

Göcke, Roland, iii, 51
Gunes, Hatice, 35

Kaiser, Robin, 105
Kuffner, Adam, 29
Kumar, Dinesh Kant, 13, 67, 93

Lucey, Patrick, 79
Lucey, Simon, 43

Matthews, Iain, 3, 43

Naik, Ganesh R., 67

Oertel, Karina, 105

Palaniswam, Marimuthu, 67
Peixoto, Paulo, 73
Piccardi, Massimo, 35
Potamianos, Gerasimos, 5, 7
Powers, David, 9

Robles-Kelly, Antonio, iii, 29

Salvador, Jordi, 19
Saragih, Jason, 51
Singh, Vijay Pal, 67
Sridharan, Sridha, 79, 87

Wagner, Michael, 107
Wark, Tim, 87
Weghorn, Hans, 13

Yau, Wai Chee, 13, 93

Recent Volumes in the CRPIT Series

ISSN 1445-1336

Listed below are some of the latest volumes published in the ACS Series *Conferences in Research and Practice in Information Technology*. The full text of most papers (in either PDF or Postscript format) is available at the series website <http://crpit.com>.

Volume 53 - Conceptual Modelling 2006

Edited by Markus Stumptner, *University of South Australia*, Sven Hartmann, *Massey University, New Zealand* and Yasushi Kiyoki, *Keio University, Japan*. January, 2006. 1-920-68235-X.

Contains the proceedings of the Third Asia-Pacific Conference on Conceptual Modelling (APCCM2006), Hobart, Tasmania, Australia, January 2006.

Volume 54 - ACSW Frontiers 2006

Edited by Rajkumar Buyya, *University of Melbourne*, Tianchi Ma, *University of Melbourne*, Rei Safavi-Naini, *University of Wollongong*, Chris Stekettee, *University of South Australia* and Willy Susilo, *University of Wollongong*. January, 2006. 1-920-68236-8.

Contains the proceedings of the Fourth Australasian Symposium on Grid Computing and e-Research (AusGrid 2006) and the Fourth Australasian Information Security Workshop (Network Security) (AISW 2006), Hobart, Tasmania, Australia, January 2006.

Volume 55 - Safety Critical Systems and Software 2005

Edited by Tony Cant, *University of Queensland*. April, 2006. 1-920-68237-6.

Contains the proceedings of the 10th Australian Workshop on Safety Related Programmable Systems, August 2005, Sydney, Australia.

Volume 56 - Vision in Human-Computer Interaction

Edited by Roland Goecke, Antonio Robles-Kelly, and Terry Caelli, *NICTA*. November, 2006. 1-920-68238-4.

Contains the proceedings of the HCSNet Workshop on the Use of Vision in Human-Computer Interaction (VisHCI 2006).

Volume 57 - Multimodal User Interaction 2005

Edited by Fang Chen and Julien Epps *National ICT Australia*. April, 2006. 1-920-68239-2.

Contains the proceedings of the NICTA-HCSNet Multimodal User Interaction Workshop 2005, Sydney, Australia, 13-14 September 2005.

Volume 58 - Advances in Ontologies 2005

Edited by Thomas Meyer, *National ICT Australia, Sydney* and Mehmet Orgun *Macquarie University*. December, 2005. 1-920-68240-6.

Contains the proceedings of the Australasian Ontology Workshop (AOW 2005), Sydney, Australia, 6 December 2005.

Volume 60 - Information Visualisation 2006

Edited by Kazuo Misue, Kozo Sugiyama and Jiro Tanaka. February, 2006. 1-920-68241-4.

Contains the proceedings of the Asia-Pacific Symposium on Information Visualization (APVIS 2006), Tokyo, Japan, February 2006.

Volume 61 - Data Mining and Analytics 2006

Edited by Peter Christen, *Australian National University*, Paul J. Kennedy, *University of Technology, Sydney*, Jiuyong Li, *University of Southern Queensland*, Simeon Simoff, *University of Technology, Sydney* and Graham Williams, *Australian Taxation Office*. December, 2006. 1-920-68242-2.

Contains the proceedings of the Australasian Data Mining Conference (AusDM 2006), Sydney, Australia. December 2006.

Volume 62 - Computer Science 2007

Edited by Gillian Dobbie, *University of Auckland, New Zealand*. January, 2007. 1-920-68243-0.

Contains the proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007), Ballarat, Victoria, Australia, January 2007.

Volume 63 - Database Technologies 2007

Edited by James Bailey, *University of Melbourne* and Alan Fekete, *University of Sydney*. January, 2007. 1-920-68244-9.

Contains the proceedings of the Eighteenth Australasian Database Conference (ADC2007), Ballarat, Victoria, Australia, January 2007.

Volume 64 - User Interfaces 2007

Edited by Wayne Epps, *University of South Australia*. January, 2007. 1-920-68245-7.

Contains the proceedings of the Eighth Australasian User Interface Conference (AUIC2007), Ballarat, Victoria, Australia, January 2007.

Volume 65 - Theory of Computing 2007

Edited by Joachim Gudmundsson, *NICTA, Australia* and Barry Jay *UTS, Australia*. January, 2007. 1-920-68246-5.

Contains the proceedings of the Thirteenth Computing: The Australasian Theory Symposium (CATS2007), Ballarat, Victoria, Australia, January 2007.

Volume 66 - Computing Education 2007

Edited by Samuel Mann, *Otago Polytechnic and Simon Newcastle University*. January, 2007. 1-920-68247-3.

Contains the proceedings of the Ninth Australasian Computing Education Conference (ACE2007), Ballarat, Victoria, Australia, January 2007.

Volume 67 - Conceptual Modelling 2007

Edited by John F. Roddick, *Flinders University* and Annika Hinze, *University of Waikato, New Zealand*. January, 2007. 1-920-68248-1.

Contains the proceedings of the Fourth Asia-Pacific Conference on Conceptual Modelling (APCCM2007), Ballarat, Victoria, Australia, January 2007.

Volume 68 - ACSW Frontiers 2007

Edited by Ljiljana Brankovic, *University of Newcastle*, Paul Coddington, *University of Adelaide*, John F. Roddick, *Flinders University*, Chris Stekettee, *University of South Australia*, Jim Warren, *University of Auckland*, and Andrew Wendelborn, *University of Adelaide*. January, 2006. 1-920-68249-X.

Contains the proceedings of the ACSW Workshops - The Australasian Information Security Workshop: Privacy Enhancing Systems (AISW), the Australasian Symposium on Grid Computing and Research (AUSGRID), and the Australasian Workshop on Health Knowledge Management and Discovery (HKMD), Ballarat, Victoria, Australia, January 2007.

Volume 72 - Advances in Ontologies 2006

Edited by Mehmet Orgun *Macquarie University* and Thomas Meyer, *National ICT Australia, Sydney*. December, 2006. 1-920-68253-8.

Contains the proceedings of the Australasian Ontology Workshop (AOW 2006), Hobart, Australia, December 2006.

Volume 73 - Intelligent Systems for Bioinformatics 2006

Edited by Mikael Boden and Timothy Bailey *University of Queensland*. December, 2006. 1-920-68254-6.

Contains the proceedings of the AI 2006 Workshop on Intelligent Systems for Bioinformatics (WISB-2006), Hobart, Australia, December 2006.